Attack Modelling: Towards a Second Generation Watermarking Benchmark

S. Voloshynovskiy, S. Pereira, V. Iquise, and T. Pun

University of Geneva - CUI, 24 rue General Dufour, CH 1211,Geneva 4, Switzerland
http://cuiwww.unige.ch/~vision/
E-mail: {svolos,Shelby.Pereira,Thierry.Pun}@cui.unige.ch

Abstract

Digital image watermarking techniques for copyright protection have become increasingly robust. The best algorithms perform well against the now standard benchmark tests included in the Stirmark package. However the stirmark tests are limited since in general they do not properly model the watermarking process and consequently are limited in their potential to removing the best watermarks. Here we propose a stochastic formulation of watermarking attacks using an estimation-based concept. The proposed attacks consist of two main stages: a) watermark or cover data estimation; b) modification of stego data aiming at disrupting the watermark detection and of resolving copyrights, taking into account the statistics of the embedded watermark and exploits features of human visual system.

In the second part of the paper we propose a "second generation benchmark". We follow the model of the Stirmark benchmark and propose the 6 following categories of tests: denoising attacks and wavelet compression, watermark copy attack, synchronization removal, denoising/compression with following perceptual remodulation, denoising and random bending. Our results indicate that even though some algorithms perform well against the Stirmark benchmark, almost all algorithms perform poorly against our benchmark. This indicates that much work remains to be done before claims about "robust" watermarks can be made. We also propose a new method of evaluating image quality based on the Watson metric which overcomes the limitations of the PSNR.

1 Introduction

Digital watermarking has emerged as an appropriate tool for the protection of author's rights [12]. It is now well accepted that an effective watermarking scheme must successfully deal with the triple requirement of *imperceptibility* (visibility) - *robustness* - *capacity* [58]. *Imperceptibility* requires that the

Preprint submitted to Elsevier Science *20 June 2001*

marked data and the original data should be perceptually undistinguishible. *Robustness* refers to the fact that the embedded information should be reliably decodable after alterations of the marked data. Often the level of robustness is dictated by the application. *Capacity* requires to the amount of information that is being embedded in the watermark. In typical applications we require between 60 and 100 bits. This is necessary so as to associate uniquely associate images with buyers and sellers. Additional requirements on the design of watermarking systems are *security*, i.e. the embedded data should be only decodable by the authorized party, and so called *blind* or *oblivious* detection and decoding, i.e. the embedded data must be decodable without referring to the original data.

Given the relatively complex tradeoffs involved in designing a watermarking system, the question of how to perform fair comparisons between different algorithms naturally arises. A lack of systematic benchmarking of existing methods however creates confusion amongst content providers and watermarking technology suppliers. Existing benchmarking tools like Stirmark [55] or Unzign [1] integrate a number of image processing operations or geometrical transformations aimed at removing watermarks from a stego image. However, the quality of the processed image is often too degraded to permit further commercial exploitation. Moreover, the design of these tools does not take into account the statistical properties of the images and watermarks in the design of attacks. As a result, pirates can design more efficient attacks that are not currently included in the benchmarking tools. This could lead to a tremendous difference between what existing benchmarks test and real world attacks.

Within this context, the goal of this article is threefold. First, we present a survey of methods which attempt to remove watermarks. Secondly, in the spirit of Fabien Petitcolas' Stirmark benchmarking tool, we propose a second generation benchmark which attacks watermarking schemes in a more effective manner. In particular the attacks contained in our benchmark take into account prior information about the image and watermark as well as the watermarking algorithm used. Our main conclusion is that while several algorithms perform well against the benchmark proposed by Petitcolas, the algorithms we evaluate almost all perform poorly relative to the proposed benchmark. This suggests that although claims about "robust" watermarks persist in the literature, the reality of the situation as demonstrated by systematic testing is otherwise. Thirdly, we propose the use of Watson's metric as a fair criteria for comparing the visibility of different watermarking schemes. We show that PSNR as proposed by Petitcolas is inadequate, and that Watson's metric is quite robust in yielding a fair comparison between algorithms.

This paper presents a general model for watermark attacks based on the above mentioned weak points of existing methods. The investigation of these weak points is performed in Sections 3 and 4 with respect to a communication formulation of digital image watermarking, which is decomposed into message embedding and extraction processes. Section 4 presents the second generation attacks which are based on the estimation concept. The attacks based on the estimate of the cover data are considered in Section 6 and corresponding attacks based on the estimate of the watermark in Section 7. Section 8 outlines the possible countermeasures against estimation-based attacks. Section 9 and 10 present the generalized attacking concept as a game between data hider and attacker. New perceptual quality metrics are considered in Section 11 and the second generation benchmarking is discussed in Section 12. Section 13 concludes the paper.

2 State-of-art Watermarking attacks

We will adopt the attack classification scheme presented in the previous paper [37]. The wide class of existing attacks can be divided into four main categories: removal attacks, geometrical attacks, cryptographic attacks and protocol attacks. Figure 1 summarizes the different attacks. We will now closely analyze each concept.

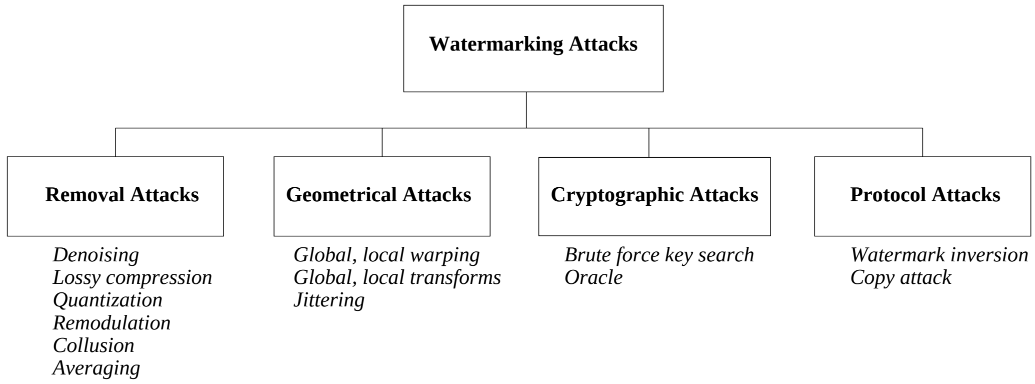


Fig. 1. Classification of watermarking attacks.

2.1 Removal attacks

Removal attacks aim at complete removal of a watermark from the cover data. This category includes denoising, lossy compression, quantization, remodulation, collusion and averaging attacks.

2.1.1 Denoising and lossy compression attacks

The basic idea of this approach consists in the assumption that the watermark is noise which can be modeled statistically. Therefore, estimating the original, non-watermarked cover data based on an available copy of the stego data, an attacker can potentially achieve the desired goal of watermark removal. This class of attacks is quite wide and includes image processing operations such as image denoising, lossy compression, quantization and remodulation. Image denoising, also known as filtering is mostly based on a Maximum Likelihood (ML), a Maximum a Posteriori Probability (MAP), a Minimum Mean Square Error (MMSE) or a minimax criteria. The resulting filtering algorithm is determined by the chosen criteria as well as by the priors on the watermark and the cover image. We will constrain our review to the ML and MAP estimates which are the most frequently encountered in practice.

In the case of the ML, the well-known denoising algorithms are local mean (or average), median, trimmed mean and myriad filter [31] which are the estimates for a Gaussian, Laplacian, $\epsilon-$ contaminated (mixture model of Gaussian and Laplacian), and Cauchy watermark distributions respectively. The averaging and median filters are included in the benchmarking tool Stirmark [56] that can be downloaded from site [55]. The representatives of the MAP-estimates are the adaptive Wiener (Lee) filter, soft- and hard- shrinkage [20] which are considered below in more details with respect to the watermarking applications. A detailed analysis of the denoising methods is given in Section 6.1. Lossy compression has recently been established to have roughly the same influence on noise removal as denoising [62,46]. This was reported with respect to watermarking applications in [69] and is considered in Section 6.2.

Both the denoising and the lossy compression can significantly reduce the capacity of watermarking channel setting to the zero the output of many equivalent channels for every bit of watermark. First of all, that is true for the flat areas where the image coefficients are assigned to zero in some transform domain without image quality degradation. The compression ratio of modern coders in wavelet domain reaches about 40-60 times with acceptable quality.

2.1.2 Remodulation attack

Since denoising and lossy compression have been extensively studied in the literature with respect to applications of image enhancement and low bit rate coding respectively, it is not surprising that they are now also well known to the watermarking community as an attack tool. On the other hand, attacks based on remodulation are a relatively new concept unique to the watermarking attack problem.

An efficient remodulation attack was first presented by Langelaar *et al* [38,39].

In this scheme the watermark was predicted via subtraction of the median filtered version of stego image from the stego image itself. The predicted watermark was additionally high-pass filtered, truncated and then subtracted from the stego image with a constant amplification factor of 2. Since median filter mostly removes the high-frequency part of noise, the low-frequency part can not be accurately estimated based on this filter. In the case of a strong match between the estimated watermark and the amplipification factor, the attack can lead to a decrease in overall correlation in the matched filter at decoding. However, as it is noticed by the authors [38,39], this scheme can perform well only for high pass watermarks. In a real scenario, when the watermark statistics/spectrum are matched with the image to guarantee a visual imperceptability, this attack shows poor performance. A similar attack based on weighted mean prediction was proposed by Holliman *at a*l [30] where authors report their success in removal of watermarks produced by the watermarking scheme proposed by Pitas [57] and the commercial software of Digimark [10].

Su and Girod [64] propose a "Wiener attack". The proposed attack consists of three steps: prediction of the watermark based on the Wiener filter, subtraction of the estimated watermark from the stego with some strength factor, and addition of stationary Gaussian noise. In the second step, the strength factor is determined based on the condition of minimization of cross-correlation coefficient between the attacked image and the watermark. The authors conclude that the addition of stationary noise does not help the attacker reach his/her goal and in practice the third step is omitted so as not to degrade image quality. In order to resist against the Wiener attack it is proposed to make the estimation of the watermark difficult for the attacker. It leads to the formulation of a power-spectrum condition that states that the watermark should look like the original in terms of power spectra to be energy-efficient. This attack has several drawbacks in the case of content adaptive watermarking when the strength of the watermark is different for different image regions. This is connected with the main assumption that the watermark as well as the image are zero-mean, wide-sense stationary Gaussian processes; clearly this assumption is satisfied neither for real images, nor for content adaptive watermarks. Consequently, it is imperative when subtracting the estimated watermark, to take into account the content adaptive nature of the embedding algorithm. Otherwise, distortions will be unnecessarily introduced.

Moulin and O'Sullivan [45] consider the influence of the attacks from the information-theoretic point of view and come to the conclusion that the additive white Gaussian noise attack can be asymptotically optimal with respect to destroying watermark when the strength of the noise is high in comparison with the energy of watermark. This attack's success results from the reduction of the watermark-to-noise ratio in the decoder through the increase of the noise variance. However, the increase of noise variance is constrained by some measure of allowable visual distortions and therefore it can not be unlimited.

Moreover, taking into account the replicated structure of the watermark and the possible gain after optimal combination of the watermark from all periods that will increase the energy of watermark, the attacker has to increase the varaince of noise. Therefore, such an attack is useless from a practical point of view. In more recent paper Moulin [43] agrees with the conclusion that the optimal attack should consist of two cascades of the MMSE estimator of the cover data and a Gaussian noise. However, it is important to note that Moulin and O'Sullivan consider only addition of the noise mutually independent with the watermark.

In order to effectively compromise between oversmoothing due to the denoising/compression and addition of infinite additive noise, Voloshynovskiy *et al* [69] independently propose a generalized two stage attack based on denoising/compression and on spatial watermark prediction using an MAP estimate of the watermark followed by perceptual remodulation to create the least favorable noise distribution for the watermark decoder. Both stages are fundamentally different from the attack proposed by Su and Girod.

With respect to the first stage, the difference with the Su and Girod attack is based on the fact that the proposed MAP estimate uses more realistic assumptions about image statistics using either a non-stationary Gaussian or stationary Generalized Gaussian model. In order to resist against this attack and to satisfy the condition of watermark imperceptibility the authors come to the same conclusion that the watermark should be matched with the statistics of the cover image. This can accomplished by adaptive watermarking where the embedding is stronger in textured regions as determined by a noise visibility function (NVF) [68]. The proposed NVF based on the non-stationary Gaussian stochastic model completely coincides with the empirically derived formula for the perception of additive noise in the different texture regions.

The second stage of the attack also contains a significant difference from the attack proposed by Su and Girod. Instead of subtracting the estimated watermark with constant strength factor, the estimated watermark is subtracted from the stego image with a local amplitude bounded by the perceptual visibility constraints. Furthermore, rather adding independent to the watermark Gaussian noise, as it is suggested by Moulin and O'Sullivan, Voloshynovskiy *at al* propose adding outliers with the sign opposite to the sign of the estimated watermark. The motivation for this is that matched filters, typically used in the watermark recovery process, are optimal with respect to Gaussian noise, however they perform poorly in general non-Gaussian noise.

2.1.3 Averaging and collusion attacks

Other attacks in this group are statistical averaging and collusion attacks. The former describes an attack in which many instances of a given data set, each time signed with a different key or different watermark, are averaged to compute the attacked data. For example, each frame can be marked using different watermark or different key in video watermarking. If the number of data sets is large enough, the embedded watermark may not be detected anymore assuming that on average it will yield zero mean. With the collusion attack, many instances of the same data are available, but this time the attacked data set is generated by taking only a small part of each data set and rebuilding a new attacked data set from these parts. Deguillaume [18] considers the averaging and collusion attacks in application to videos and proposes corresponding countermeasures.

The other type of attack that impairs the detection and decoding of the watermark is mosaic attack [56]. This attack was created within the framework of an automatic copyright protection systems that scan Internet and download images to check the presence of the watermarked images on pirate sites. The mosaic attack does not try to remove the watermark using some signal processing methods, but rather it aims at creation of the problems for the watermark detector dividing image on the small fragments. The fragments are then presented on the site as a whole image in a HTML page. Thus if the fragment is small enough to contain the complete period of the watermark the detector fails to detect it. In order to avoid this attack the watermarking methods should be robust enough to allow decoding of the watermark from very small images. This requirement is even more strict than cropping since the commercial quality of the image is preserved. We can also predict that a more intelligent system will stick back the small fragments of the images to form a bigger image and to check the presence of the watermark. We call it "screen shot" detection.

2.2 Geometrical attacks

In contrast to the removal attacks, geometrical attacks intend not to remove the embedded watermark itself, but to distort it through spatial or temporal alterations of the stego data. The attacks are usually such that the watermark detector loses synchronization with the embedded information. The most well know integrated software versions of these attacks are Unzign and Stirmark. Unzign [1] introduces local pixel jittering and is very efficient in attacking spatial domain watermarking schemes. Stirmark [55] introduces both global geometrical and local distortions. The global distortions are rotation, scaling, change of aspect ratio, translation and shearing that belong to a class of general affine transformations. The line/column removal and cropping/translation are also integrated in Stirmark. Most recent watermarking methods survive after these attacks due to the usage of special synchronization techniques. Robustness to the global geometrical distortions rely on the use of either a transform invariant domain [48], or an additional template [50,10,51], or an Autocorrelation Function (ACF) of the watermark itself [35,67].

If robustness to global affine transformations is more or less a solved issue, the local random alterations integrated in Stirmark still remains an open problem almost for all techniques. The so called random bending attack exploits the fact that the human visual system is not sensitive against shifts and local affine modifications. Therefore, pixels are locally shifted, scaled and rotated without significant visual distortions. In Section 7.3, we will also consider an new dedicated attack which aims at removing global synchronization of the above considered methods.

2.3 Cryptographic attacks

Cryptographic attacks are very similar to the attacks used in cryptography. There are the brute force attacks which aim at finding secret information through an exhaustive search. Since many watermarking schemes use a secret key it is very important to use keys with a secure length. Another attack in this category is the so called Oracle attack [54,14] which can be used to create a non-watermarked image when a watermark detector device is available.

2.4 Protocol attacks

The protocol attacks aim at attacking the concept of the watermarking application. The first protocol attack was proposed by Craver et al [16]. They introduce the framework of invertible watermark and show that for copyright protection applications watermarks need to be non-invertible. The idea of inversion consists of the fact that an attacker who has a copy of the stego data can claim that the data contains also the attacker's watermark by *subtracting* his own watermark. This can create the situation of ambiguity in respect to the real ownership of the data. The requirement of non-invertability on the watermarking technology implies that it should not be possible to extract a watermark from non-watermarked image. As a solution to this problem, the authors propose to make watermarks signal-dependent by using a one-way function.

The copy attack [37] belongs to the last group of the protocol attacks. The goal of the attack is not to destroy the watermark or impair its detection, but consists rather in the prediction of the watermark from the cover image, like in the case of the remodulation attack, followed by copying the predicted watermark on the target data. The estimated watermark is then adapted to the local features of the stego data to satisfy its imperceptibility. The process of copying the watermark requires neither algorithmic knowledge of the watermarking technology nor the watermarking key. However, in the published version of this attack it was assumed that the watermarking algorithm exploits linear additive techniques. The derivation of the optimal MAP estimate for multiplicative watermarks or generally non-additive techniques is required to cover methods like SysCop of MediaSec [42], Barni [7] and Pereira [52,53] that are mostly used in the transform domains.

Although the above classification makes it possible to have a clear separation between the different classes of attacks, it is necessary to note that very often a malicious attacker applies not only a single attack at the moment, but rather a combination of two or more attacks. Such a possibility is predicted in the Stirmark benchmark where practically all geometrical transformations are accompanied by lossy compression.

3 Problem formulation: modern digital watermarking paradigm

Having reviewed the state of the art with respect to watermarking attacks, we now model the basic concept of linear embedding algorithms. This model will be the basis for deriving new dedicated attacks which will be included in the proposed second generation of benchmarking tools.

Consider the general model of a watermarking system according to a communications formulation. Its block diagram is shown in Figure 2. The watermarking

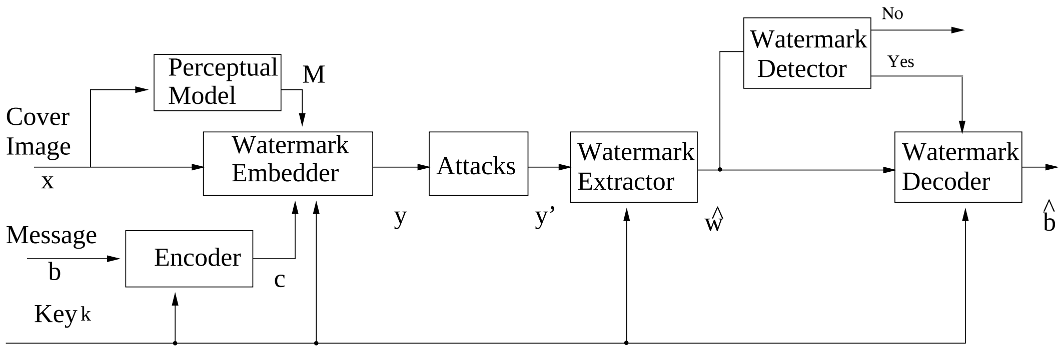


Fig. 2. Communication formulation of a watermarking system.

system consists of three main parts, i.e. message embedding, attack channel and message extraction. Let us consider in details these main parts and the corresponding weaknesses that could be used by an attacker.

3.1 Message embedding

A message $b = (b_1, ..., b_L)$ is to be embedded in the cover image $x = (x_1, ..., x_N)^T$ of size $M_1 \times M_2$, where $N = M_1 \cdot M_2$. The message b contains information about the owner and can be used for authentication purposes. To convert the message into a form efficient for communication, it is encoded using either error correction codes (ECC) or modulated using binary antipodal signaling [29] or M-ary modulation [35]. With respect to ECC, mostly Bose Chaudhuri (BCH) or convolutional codes are used [48,28]. Recent publications [52,67] report the successful results using novel Turbo codes and low-density parity-check (LDPC) codes in the DCT and wavelet domains. In the general case, the type of ECC and the set of basis functions for M-ary modulation can be key-dependent. The above conversion is performed in the encoder that produces the codewords $c = Enc(b, Key)$, $c = (c_1, ..., c_K)^T$ which are mapped from {0,1} to {-1,1} using binary phase shift keying (BPSK) modulation.

A watermark w is created by some key-dependent function $w = \varepsilon\left(c, p, M, Key\right)$ that ensures the necessary spatial allocation of the watermark based on a key-dependent projection function p, and according to HVS features as expressed by a perceptual mask M in order to improve the watermark. The typical choice for the projection function p is a set of two dimensional orthogonal functions used for every codeword bit $\{c_k\}$ such that the empty set is formed by the intersection $P_k \cap P_l, \forall k \neq l$ [35,29,4]. The projection function performs a "spreading" of the data over the image area. It can be also considered as diversity communication problem with parallel channels. Moreover, the projection function can have a particular spatial structure with given correlation properties that can be used for the recovery of affine geometrical transformations [34,67]. The resulting watermark is obtained as the superposition

$$w\left(j\right) = \sum_{k=1}^{K} c_k p_k\left(j\right) M\left(j\right) \tag{1}$$

where $j \in Z$. The watermark embedder performs the insertion of the watermark into the cover image in some transform or coordinate domain, yielding the stego image:

$$y = T^{-1}\left[h\left(T\left[x\right], w\right)\right] \tag{2}$$

where T is any orthogonal transform like block DCT, full-frame FFT and DCT, wavelet or Radon transforms ($T = I$ for the coordinate domain), and $h\left(.,.\right)$ denotes the embedding function. The most widely used class of embedding functions conforms to the linear additive model

$$y = h\left(x, w|M\right) = x + w(M) \tag{3}$$

that is considered in this paper.

To extend the above model in the more general formulation, one can consider the watermarking as a communication with side information (SI) as it is shown in Figure 3.

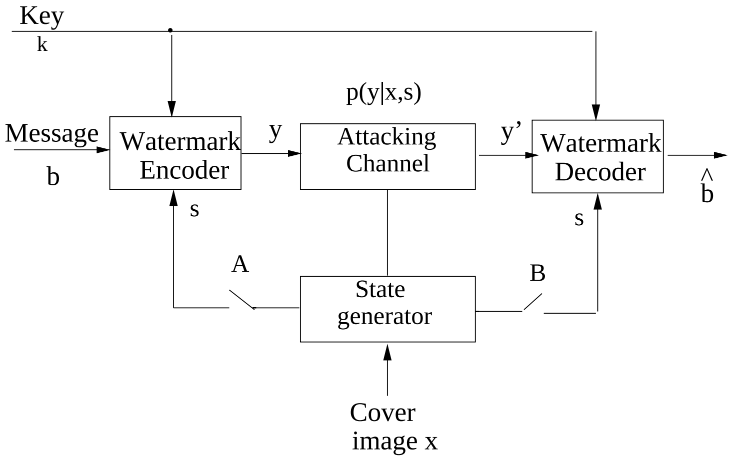


Fig. 3. Watermarking as communication with side information.

Let us consider the side information in watermarking applications assuming that the encoder and decoder have the access at least to: First, both the encoder and the decoder can access the key used for the watermark embedding. Second, the generalized channel state information can be available. The generalized channel includes the cover data and the attacking channel. The attacking channel includes all possible intentional or unintentioanl attacks that can be applied during the "life cycle" of the image or videos.

Depending on different combinations of the switches A and B in Figure 3, all watermarking algorithms can be divided into the four classes described below. It is assumed that the switch A controls the access of the encoder to the channel state information given by the cover image. The switch B controls the access of the decoder to the attacking channel information given by all possible signal processing and geometrical attacks. The key k is assumed to be available for both encoder and decoder in private watermarking.

Class I: SI is not available (switches A and B open). It is a typical case for all earlier watermarking algorithms that were inspired by the original papers of Cox [13] and Tirkel [65]. It assumes that the watermark is embedded in the cover image and is then decoded without reference to information about channel state. The detection of watermarks is mostly based on the direct correlation of the stego data with the watermark generated based on the key. If the correlation coefficient is above some threshold, then the decision is made of successful detection. As a result, the performance of these schemes is very poor due to two basic assumptions made: all attacks are modeled as additive stationary Gaussian noise that results in the simple correlation detection receiver [59,24]. Secondly, these schemes assume no geometrical attacks.

Class II: SI is available at encoder only (A open, B closed). This scheme has found recently a lot of attention in the watermarking community due to the publication of Cox [15]. The block diagram of this scheme is shown in Figure 4.

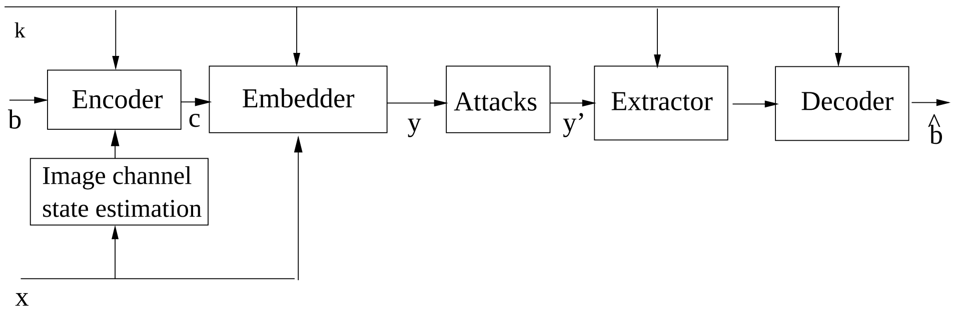


Fig. 4. Watermarking as communication with side information regarding cover data available for the encoder.

The main idea of watermarking as communication with SI at the encoder consists of the fact that the theoretical capacity of oblivious watermarking scheme is equal to that of a decoder with access to the cover data. This conclusion is based on the remarkable paper of Costa [11]. Therefore, there is no more need in the cover data for the decoder, if the cover data is used as SI by the encoder. However, this approach has several drawbacks. First, the complexity of the encoder is very high. This means that the codebook for every particular image becomes quite large. Therefore, the decoder should also perform a quite complex search. Second, the watermarking channel is only treated as the cover image and attacks are not taken into account that can lead to the missmatch between what was assumed for the design of the codebook and the real situation.

To reduce the complexity of the encoder, different practical algorithms were proposed [9,21,52]. However, geometrical attacks and the attacks whose statistics are different from additive Gaussian remain an open issue.

To relax the lack of the decoder adaptivity to the attacking channel state one can include the worst case attack as information about channel state in the decoder assuming that the decoding will be successful in more favorable conditions. An example of this approach aiming at resisting against low lossy JPEG compression is proposed by Pereira *et al*[52]. The JPEG quantization table for the worst case of Stirmark compression at quality factor QF=10 was included in the design of the encoder. As the result, the decoder can detect watermark even from very small block of size 64x64 after the above compression.

Class III: SI is available at decoder only (A closed, B open). These schemes are able to estimate the undergone attacks in the attacking channel and are potentially able to resist against geometrical trasformations. This relies on the fact the a key-dependent pilot or reference watermark can be used for two purposes. First, the pilot can be considered as the the synchronization pattern in some coordinate or transform domain, i.e. mostly in the magnitude spectrum of the DFT due to the known shift and cropping invariant properties, as well as with the simultaneous ability to detect affine transforms [34,51,67]. Secondly, the pilot embedded in the stego data can be used to estimate fading due to data embedding and attacks, and statistics of noise, if they are different from Gaussian as is a case with a lossy JPEG compression attack. This enables to consider the watermarking as a channel with fading and non-Gaussian noise and leads to diversity reception since the watermark is replicated over the image area. The pilot can be easily regenerated in the decoder based on the key.

Class IV: SI is available at both encoder and decoder (A open, B open). This scenario can be considered as the most likely scheme for all future watermarking algorithms that can operate under a wide class of uncertanties with respect to the channel state. The optimality of this scheme is based on the optimal design of the encoder matched with the cover data and adaptivity of the decoder to the attacking channel state assuming fading, non-Gaussian attacks and geometrical transforms utilizing advantages of diversity watermarking. The generalized block diagram of this scheme is shown in Figure 5.

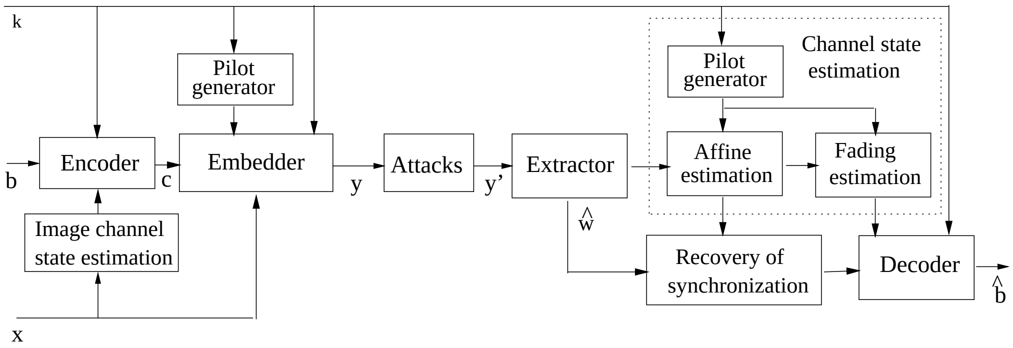


Fig. 5. Watermarking as the diversity comminucation with the side information about cover data.

3.2 *Attacking channel*

An attacking channel produces the distorted version y' of the stego image y. The attacking channel can be modeled in the framework of stochastic formulation using a probability mass function (p.m.f) $Q(y'|y)$ to describe random distortions in the stego image. A successful attack should damage or destroy the watermark while preserving the commercial quality of the image. Therefore, an attacker should introduce distortions that are value by some upper allowable bound according to the chosen distortion criterion. Although, the

MSE is not perfectly matched with the subjective human assessment of image quality, it is commonly used due to the obtained tractable results, and the wide usage of this criteria in the communication community due to the known results for the additive Gaussian channels. Therefore, the aim of the attacker consists in decreasing of the rate of reliable communication subject to the allowable distortion.

However, it is necessary to note that the above consideration will not be complete without geometrical attacks. The geometrical attacks can be mathematically modeled as affine transforms with some random parameters that are not known for the decoder. Normally, there are 6 parameters that produce all set of global affine geometrical alterations: scaling, change of aspect ratio, shearing, rotation and shift. More generally, these modifications can be modeled as projective transformations that can occur in the applications such as the "Internet bridge" of Digimark, i.e. reading the watermark in front of a web camera. The random local distortions integrated in the Stirmark benchmark and also known as random bending attack can be presented as local affine transform. Therefore, the decoder should access these parameters to have an optimally synchronized watermark decoding. The concept of pilot or reference watermark considered above might be an appropriate solution of this problem both for the affine parameters estimation and for estimation of $Q\left(y'|y\right)$ parameters.

3.3 Message extraction

The recovery process consists of the watermark extractor and decoder which are described below.

3.3.1 Watermark extractor for oblivious watermarking

The watermark extractor performs an estimate $\widehat{w}$ of the watermark based on the attacked version $\widehat{y}$ of the stego-image:

$$\widehat{w} = Extr\left(T\left[y'\right], Key\right) \tag{4}$$

In the general case, the extraction should be key-dependent. However, the desire to recover data after affine transformation based on the above mentioned self-reference principle, and the opportunity to enhance the decoding performance by reducing the variance of the image considered as noise [35,27], have motivated the development of key-independent watermark extraction methods. They could represent the main danger to linear additive watermarking technologies, as will be shown below.

Different methods are used for watermark estimation, such as the cross-shaped filter [34], or MMSE estimates [29]. In the most general case, the problem of watermark estimation can be solved based on a stochastic framework by using Maximum Likelihood (ML), penalized ML or MAP estimates [68]. Assuming that both the noise due to the cover image and the noise introduced by an attack can be considered additive with some target distribution $p_X(.)$, one can determine the ML-estimate:

$$\widehat{w} = \arg\max_{\widetilde{w}\in\Re^N} p_X(y' \mid \widetilde{w}) \tag{5}$$

which results either in a local average predictor/estimator in the case of a locally stationary independent identically distributed (i.i.d.) Gaussian model of $p_X(.)$, or a median predictor in case of a corresponding Laplacian p.d.f.. If there is some prior information about watermark statistics, the MAP estimate can be used:

$$\widehat{w} = \arg\max_{\widetilde{w}\in\Re^N}\{ p_X(y' \mid \widetilde{w}) \cdot p_W(\widetilde{w})\} \tag{6}$$

where $p_W(.)$ is the p.d.f. of the watermark. The difference between the ML and the MAP estimates consists in the fact that the MAP estimate reduces to the ML estimate, if there is no prior about watermark distribution or this prior is uniform for the set of possible solutions. To solve problems (5) and (6) it is necessary to develop accurate stochastic models for the cover image $p_X(x)$ and the watermark $p_W(w)$.

3.3.2 Stochastic models of cover image: source generation

Stochastic models of cover image applied to content adaptive watermarking were considered in our previous work [68]. We use here the main results of this work and consider either locally i.i.d. non-stationary Gaussian (nG) or globally i.i.d. Generalized Gaussian (sGG) image models. The motivation for these two models are their wide usage in a number of image processing applications including image denoising, restoration and compression, and the existence of tractable closed form solutions of (21) for the particular cases of these models.

The non-stationary Gaussian model is characterized by a distribution:

$$p_x(x) = \frac{1}{(2\pi)^{\frac{N}{2}}} \cdot \frac{1}{|\det R_x|^{\frac{1}{2}}} \cdot \exp\{-\frac{1}{2}(Cx)^T R_x^{-1} Cx\}, \tag{7}$$

where R_x is covariance matrix, $|\det R_x|$ denotes the matrix determinant, and Cx represents a high-pass filtering (decomposition operator) and it can be also rewritten as $Cx = (I - A)x = x - Ax = x - \bar{x}$, where I is the unitary matrix, A is a low-pass filter used to compute the non-stationary local mean $\bar{x}$. C could also be considered as wavelet decomposition operator in which case model (7) is used for every subband.

The stationary GG model has stationary R_x and can be written as:

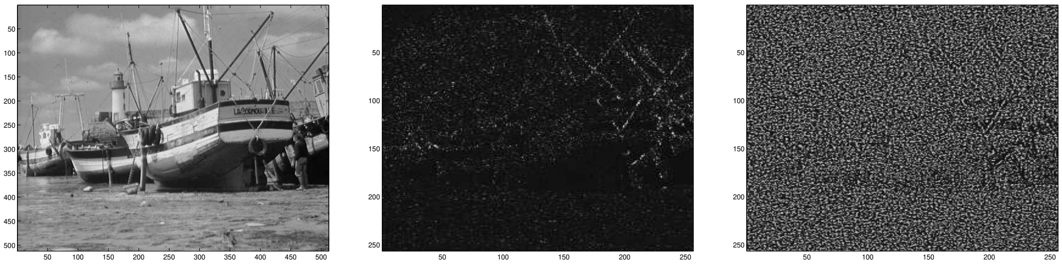
$$p_x(x) = (\frac{\gamma\eta(\gamma)}{2\Gamma(\frac{1}{\gamma})})^{\frac{N}{2}} \cdot \frac{1}{|\det R_x|^{\frac{1}{2}}} \cdot \exp\{-\eta(\gamma)(|Cx|^{\frac{\gamma}{2}})^T R_x^{-\frac{\gamma}{2}} |Cx|^{\frac{\gamma}{2}}\}, \qquad (8)$$

where $\eta(\gamma) = \sqrt{\frac{\Gamma(\frac{3}{\gamma})}{\Gamma(\frac{1}{\gamma})}}$ and $\Gamma(t) = \int_0^\infty e^{-u}u^{t-1}du$ is the gamma function, and the parameter γ is called the *shape parameter*. Equation (8) includes the Gaussian ($\gamma = 2$) and the Laplacian ($\gamma = 1$) models as special cases. For real images the shape parameter is in the range $0.3 \leq \gamma \leq 1$. The other examples of stationary stochastic models are mixture models that either include two additive Gaussian distributions with different variances, i.e. the increased variance is used to model havy tails in the distribution, or include Gaussian and Laplacian p.d.f.s. Cauchy distribution can be also used to approximate the heavy tail statistics.

There is a strict connection between local non-stationary Gaussian and global stationary Generalized Gaussian models. If we consider the image locally, then it could be accurately modeled by the non-stationary Gaussian model, while treating the same data globally as i.i.d with the same variance one can approximate it using stationary GG model for a particular γ. To show this connection we will consider an example in the wavelet domain; the consideration is also valid for coordinate domain modeling and for multichannel DCT based image representations used in current JPEG compression standard.

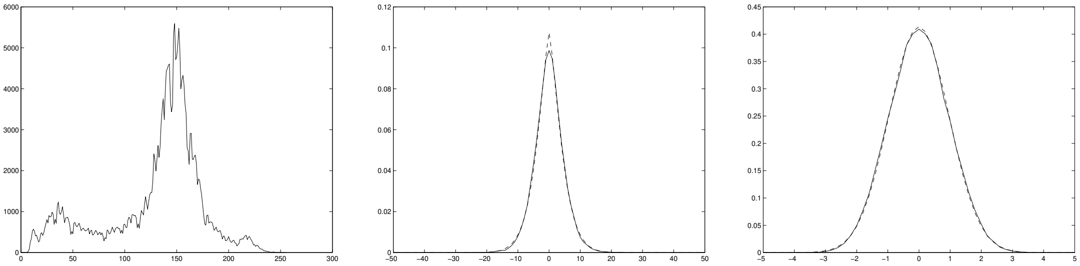
The original Boat image (Figure 6a) is decomposed using a wavelet transform. The first scale coefficients for the diagonal orientation sub-band are shown in Figure 6b. It is necessary to note that the same results can be obtain using a Laplacian image decomposition pyramid or simply by subtracting the local image mean estimated in a window of size 5x5 that will aproximate the Laplacian operator. The above image has non-stationary character, i.e. the regions of edges and textures are more visible and have larger amplitude due to the edge transitions. To normalize the image, i.e. to make its distribution close to normal or Gaussian ($N(0,1)$), we divide it by the estimate of the local standard deviation; this results in the image shown in Figure 6c, which has a more uniform character. Assuming that the image coefficients are stationary, i.e. originate from the same distribution, we plot the corresponding histogram of the images from Figure 6 that are depicted in Figure 7. It is possible to follow the changes in the histograms statistics starting from multimodal (Figure 7a) that can be modeled as a mixture of Gaussian, to unimodals (Figure 7b and c) that are quite accurately approximated using stationary Generalized Gaussian and stationary Gaussian models, respectively.

This simple experiment makes possible to establish the practically important dependencies between different stochastic models and to formulate a uniform



(a) (b) (c)

Fig. 6. a) The original Boat image b) the result of decomposition c) the normilized decompozed image.



(a) (b) (c)

Fig. 7. The histograms of: a) the original Boat image, b) the decomposed and c) the normilized images and their approximation by stationary Generalized Gaussian $sGG(0, 0.6, 17)$ and zero-mean unit variance Gaussian models $N(0, 1)$.

stochastic framework for image modeling. Based on that, the image can be treated as a multichannel stochastic process. Using inverse order of the above decomposition of the image into zero-mean unit variance Gaussian noise this multichannel model can be presented as in Figure 8. First, each pixel of the image is modeled as a stationary source with $N(0, 1)$. Secondly, each pixel is multiplied by the non-stationary standard deviation and biased by the non-stationary mean, resulting in the final observed image. Therefore, considering each pixel locally it can be presented as non-stationary mean non-stationary variance Gaussian model. At the same time treating all coefficients globally, i.e. originating from the same i.i.d. source that is represented as a multiplexor in Figure 8, one can use stationary Generalized Gaussian approximation (Figure 7b). The connection between stationary Generalized Gaussian and non-stationary Gaussian models will be further widely used in the paper for the design of optimal watermark extraction strategy, watermarking decoder and also for the derivation of optimal attacks based on the estimation-based concept and possible countermeasures. Obviously, more complex models can be used that do not have the limitations of i.i.d. models and that take into account local correlation between image pixels.

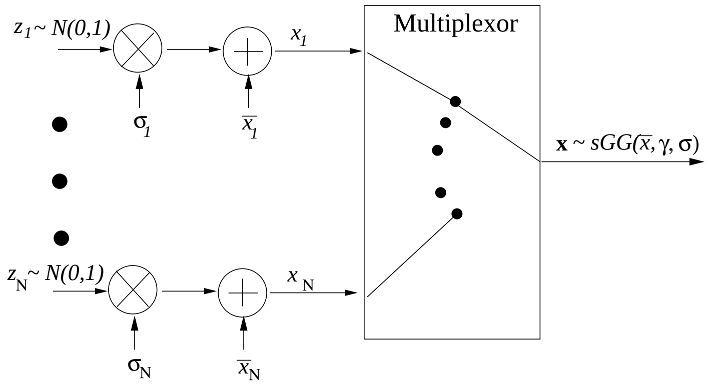


Fig. 8. The generalized multichannel stochastic model of image generation.

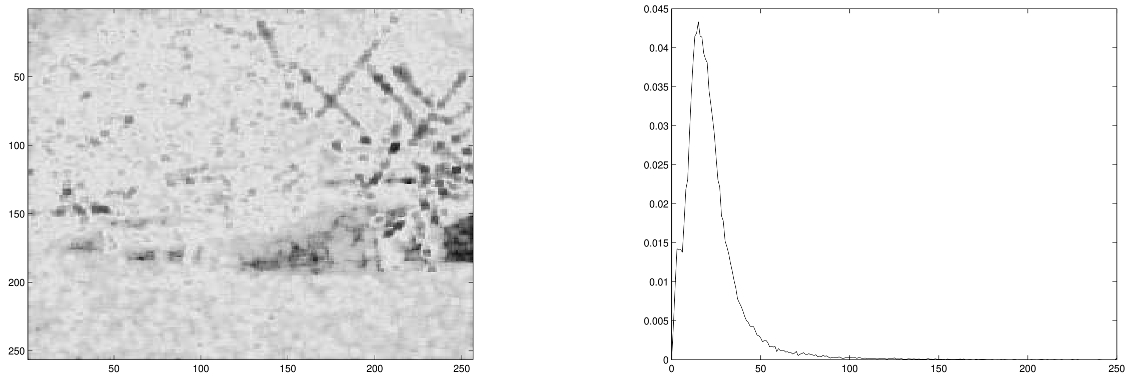
The important moment of the stochastic image modeling is the estimation of the hyperparameters of the models. In the case of the nG model one must estimate the local mean and the local variance while in the case of the sGG model the local mean, the shape parameter and the global variance should be estimated. To estimate the local image variance the *maximum likelihood* estimate can be used. Assuming that the image is a locally i.i.d. Gaussian distributed random variable, the ML estimate is given by:

$$\sigma_x^2(i,j) = \frac{1}{(2L+1)^2} \sum_{k=-L}^{L} \sum_{l=-L}^{L} (x(i+k,j+l) - \bar{x}(i,j))^2 \tag{9}$$

with

$$\bar{x}(i,j) = \frac{1}{(2L+1)^2} \sum_{k=-L}^{L} \sum_{l=-L}^{L} x(i+k,j+l), \tag{10}$$

where a window of size $(2L+1) \times (2L+1)$ is used for the estimation. This estimate is often used in practice in many applications. However, the above estimate is assymptotically unbiased. To decrease the bias, it is necessary to enlarge the sampling space. From the other side, enlarging the window size violates the requirement of data being locally Gaussian, since the pixels from different regions occur in the same local window. In order to have a more accurate model, it is reasonable to assume that flat regions have a Gaussian distribution while textured areas and regions containing edges have some other highly-peaked, near-zero distribution (for example Laplacian). An example of such estimation is shown in Figure 9. It is necessary to note that the histogram of the local variance can be approximated as Weibull or Rice distributions. Knowing the statistics of the hyperparameters, like in the case with the local variance, one can model images as doubly stochastic processes.



(a) (b)

Fig. 9. a) The local variance of the decomposed image b) corresponding histogram.

3.3.3 Stochastic model of watermark

In the general case, we can use the same models for perceptually embedded watermark based on equation (1) as for the cover image. If the used perceptual model is known for an attacker and the information about the watermark embedding method is available, one can estimate the watermark directly from the stego image as was discussed above. If there is some ambiguity with respect to these priors, a robust M-estimation approach can be used [66].

Assuming that the image and watermark are conditionally i.i.d. locally Gaussian, i.e. $x \sim N(\overline{x}, R_x)$ and $w \sim N(0, R_w)$ with covariance matrices R_x and R_w, where R_w also includes the effect of perceptual watermark modulation, one can determine:

$$\widehat{w} = (R_w + R_x)^{-1} R_w (y' - \overline{y}') \tag{11}$$

where it is assumed $\overline{y}' \approx \overline{x}$ and $\overline{y}'$ is a local mean of the attacked stego image that can be estimated based on local avarage, and where $\widehat{R}_x = \max\left(0, \widehat{R}_y - R_w\right)$ is the ML estimate of the local image variance $\left(\widehat{R}_x = \sigma_x^2 I\right)$. It is necessary to note that the local mean of the attacked image can be assumed to be zero, if the above prediction is performed in the wavelet domain. Then, the autocavariance function can be estimated using the ML estimate for every wavelet sub-band coefficient. The equation (11) is the MAP/MMSE watermark extractor for oblivious watermarking for the considered above stochastic models.

3.3.4 Watermark decoding

In the general case the decoder/demodulator design is based on ML or MAP approaches. Since the appearance of b is assumed to be equiprobable and due to the high complexity of the MAP decoders, ML decoders are mostly used in practice. The watermark decoder can be considered to consist of two main parts: a matched filter (detector) that performs a despreading of the data in the way of "coherent accumulation" of the sequence c spread in the watermark w, and the decoder itself that produces the estimate of the message. In most cases the results of attacks and of prediction/extraction errors are assumed to be additive Gaussian. The detector is therefore designed using an ML formulation for the detection of a known signal (projection sets are known due to the key) in Gaussian noise, that results in a correlator detector with reduced dimensionality:

$$r = \langle \widehat{w}, p \rangle . \tag{12}$$

Unfortunatly, the above matched filter does not take into account the practically important cases of fading and non-Gaussian noise. The equation (12) is typical for class I watermarking systems considered above. Therefore, to design a more realistic model of equivalent watermarking channel we consider the transmission of the codeword c containing the encoded watermark and pilot bits through the parallel channel according to the diversity communication (Figure 10). The parameters of the channel are estimated using the pilot. We assume that the parameters of the affine transform $\widehat{A}$ are estimated and recovered prior to the decoding.

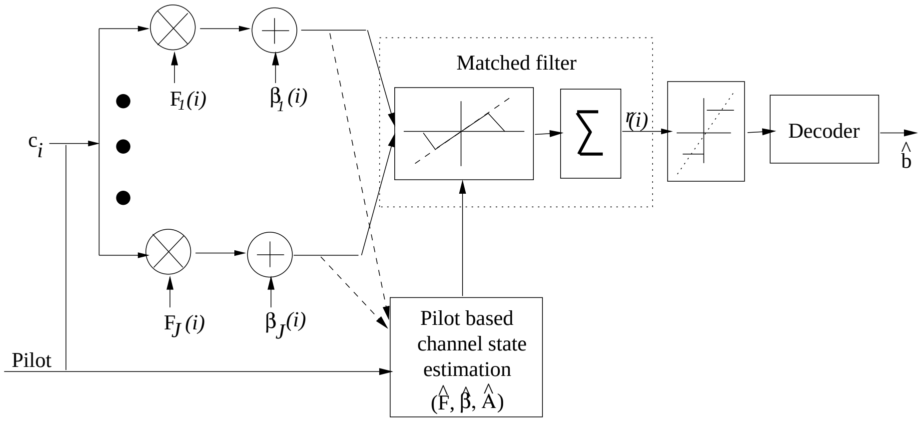


Fig. 10. Equivalent parallel channel formulation of digital watermarking.

The equivalent channel model can be presented as:

$$c' = Fc + \beta \tag{13}$$

where F denotes the generalized fading and β is generalized noise in the equivalent parallel channel. The generalized fading includes several factors. First, the watermark is masked by the perceptual mask used for the embedding that has different values for the the different periods of the watermark replication. Secondly, attacks like denoising and lossy compression significantly decrease the strength of the watermark especially in the flat image regions, even reducing it to zero. As it will be shown below, the optimal attack can perform watermark removal based on the MAP or the MMSE criteria. Thirdly, the watermark extractor also modifies the amplitude of the watermark according to the local image statistics. The generalized noise includes all possible modifications of the watermark after attack that can be described using either non-stationary Gaussian or stationary Generalized Gaussian models. In the more general case a generalized matched filter can be designed that produces the output:

$$r = \langle g(\widehat{w}), p \rangle \tag{14}$$

where $g(.)$ is determined by the statistics of the generalized channel. In the particular case of the non-stationary Gaussian noise model the matched filter will have the next structure:

$$r_{nG} = \left\langle \widehat{R}_{\beta}^{-1} \widehat{F} \widehat{w}, p \right\rangle \tag{15}$$

or equivalently

$$r_{nG}(i) = \sum_{j \in P_i} \frac{\widehat{F}(j)\widehat{w}(j)p(j)}{\sigma_{\beta}(j)^2} \tag{16}$$

for all $i = 1, ..., K$ and where j is the index of diversity or the number of replication of bit c_i, $\widehat{R}_{\beta}$ is an estimate of covariance matrix of non-stationary Gaussian noise $N(0, {\sigma_{\beta}}^2 I)$ and $\widehat{F}$ is an estimate of the channel fading. Physically, the estimation of channel parameters is performed based on the assumption that the pilot bits are closely allocated to the corresponding equivalent channel bits of the codeword in the stego image. This does not exactly correpond to the communication channel and this analogy is slightly artificial here. However, assuming some certain degree of correlation between neightborhood pixels in the image we can assume that the pilot bits will have about the same modifications as the bits of the codeword. Therefore, it can be modeled as a slow fading channel with respect to the pilot signal. However, since the codewords are allocated with some random locations over the image that can be quite remote where the local image correlation does not play any significant role, it can be modeled as fast fading channel.

It is important to note that the matched filter produces a soft output that can be important for further decoding. The scheme with the hard output assuming binary symmetric channels was first proposed by Kundur *et al* in watermarking applications [33]. This scheme is considerably simplified and does not require the estimation of the channel non-stationary variances and parameters of the fading. It is modeled using only error probablities for each channel. As a consequence, weighted coefficients are derived for the diversity summation like in (16).

The assumption about stationary Generalized Gaussian noise distribution $sgg(0, \beta, \sigma_\beta)$ leads to the following matched filter:

$$r_{sGG}(i) = \sum_{j \in P_i} \frac{\left|\widehat{w}(j) + \widehat{F}(j)p(j)\right|^\beta - \left|\widehat{w}(j) - \widehat{F}(j)p(j)\right|^\beta}{\sigma_\beta(i)^\beta} \tag{17}$$

where $\sigma_\beta(i)$ is constant for the given codeword bit c_i. The nonlinear structure of the matched filter is similar to local optimum detector nonlinearity that limits the outliers of the sGG model [32]. This model was considered for the DCT domain watermarking by Hernández *et al* [26] where the sGG model presented the distribution of the DCT coefficients in 64 equivalent channels of JPEG compression. The authors considered this model assuming that all fading in the equivalent channel is only due to the perceptual masking and the used mask was proposed to be estimated directly from the attacked stego image. Therefore, in this formulation the matched filter is not completely adopted to the channel state variations, in contrast with the above considered pilot based technique. It is important to note that both considered matched filters (16) and (17) can be applied for the coordinate, wavelet and DCT domains.

The output of the matched filter is thresholded according to either the hard or the soft decoding (Figure 10) and then decoded. A decoder can be designed based on the MAP:

$$\widehat{b} = \arg\max\nolimits_{\widetilde{b}}\, p\left(\widetilde{b} \mid r, x, k\right) \tag{18}$$

Assuming that all codewords b are equiprobable, given an observation vector r, the optimum decoder that minimizes the conditional probability of error is given by the ML decoder:

$$\widehat{b} = \arg\max\nolimits_{\widetilde{b}}\, p\left(r \mid \widetilde{b}, x, k\right). \tag{19}$$

Based on the central limit theorem (CLT) most researchers assume that the observed vector r can be accurately approximated as the output of an additive Gaussian channel noise [35,27] that can be exploited by the attacker as it will be shown below. Although the considered schemes are much more advanced in comparison with the linear correlation receiver, we will further concentrate the analysis of the attacks on the scheme (12) since it is used in the dominating majority of the existing watermarking algorithms [58–60,24,34,27,12,4,65].

Therefore, it can be a quite attractive domain for the attackers.

4 Watermark attacks based on the weak points of linear methods

A key-independent watermark prediction according to (11) presents several problems. The first problem is connected with the assumption that the stego image is not significantly altered after attack. This allows the perceptual mask used at embedding to be estimated from the attacked stego image [35,27,19]. However this assumption does not hold for attacks connected with histogram modification that could have a significant influence on models based on luminance masking, and lossy JPEG compression attack whose strong blocking artifacts could alter models based on texture masking.

Another series of problems are tied to the general security-robustness issue. Since the watermark can be predicted based on (11) without knowledge of the key, the following problems appear:

(1) The redundancy in the watermark and global watermark energy can be considerably reduced as a result of denoising and compression, this especially in flat image regions.
(2) Special types of distortions could be introduced in the watermark, aiming to create the least favorable conditions for the decoder. In particular, perceptual remodulation of the watermark aimed at creating the least favorable statistics for the AWGN decoder designed based on (12,19) will be shown to be an extremely effective attack.
(3) The synchronization can be destroyed by estimating template or the parameters of periodical watermarks and then removing the synchronization mechanism.
(4) If we ignore perceptual masking, most algorithms generate watermarks independently from the image. This leads to vulnerability with respect to the watermark copy attack in which the watermark is estimated from one image and added to another one in order to generate a falsely watermarked image.

We will consider each of these points in detail in the text that follows.

5 The second generation attacks

The attacks included in benchmarking tools such as Stirmark and Unzign are commonly used in practical image processing applications. These attacks are accessible to many inexperienced attackers and can be found in most modern image processing tools like Photo Shop or Paint Shop Pro. However, an essential drawback of these attacks is their outdated character since many watermarking technologies already integrate efficient anti-attacking tricks. We refer to these sort of attacks as the first generation attacks. The second generation attacks take into account the knowledge of watermarking technology and exploit statistics of images and watermarks to design successful attack while preserving or even enhancing image quality.

In the scope of the second generation watermarking attacks, we present the concept of estimation-based attacks. This concept is based on the assumption that the cover image or the watermark can be estimated from the stego data using some prior knowledge of the stego image and watermark statistics. It is necessary to note that the estimation does not require any knowledge of the key used for watermark embedding. The knowledge of the embedding rule is not required, but the additional gain in the success of the attack can be obtained, when the embedding rule is known.

We consider some generalized embedding rule (3). In practice, there are three mostly used embedding schemes: additive linear, multiplicative [5,6,59] and quantization index modulation (QIM) [9]. More generally, all of them could be considered from the point of view of some generalized linear additive scheme:

$$y = x + H\left(x, w|M\right) \tag{20}$$

where $H(.,.)$ is some possibly nonlinear transform.

According to the final purpose of the applied attack, the attacker can obtain the estimate of the cover data or the watermark based on some stochastic criteria such as the ML or the MAP, or the MMSE. We will not focus here on the particularities of the above estimation, but rather concentrate on the analysis and the applications of the obtained estimates.

6 Attacks based on estimate of the cover data

Considering the watermark as noise in the stego data, the attacker can try to estimate the original, unwatermarked data. The final attack will result in the design of the optimal denoising scheme. Taking into account the results of recent investigation that established the strong connection between denoising and compression for filtering of additive noise from the images, the attacker can easily apply the most recent advanced wavelet coders to remove the watermark. Keeping in mind the optimal design of such coders that are based on rate-distortion theory, the attacker can obtain a considerable gain in resolving the compromise between distortions introduced by this attack and removal of the watermark. This can also kill the QIM schemes due to the requantization. It is necessary to note that in both cases of denoising and of the above optimized compression both visual and objective quality of the attacked image can be improved by several decibels. We will refer to such attacks as the group of removal attacks.

6.1 Watermark removal based on denoising

The watermark can be removed from the stego image in some cases or its energy can be considerably decreased using denoising/compression attack. Consider the MAP estimation of the cover image as image denoising according to the additive model (3)

$$\hat{x} = \arg\max_{\tilde{x}\in\Re^N} \left\{\ln p_W\left(y \mid \tilde{x}\right) + \ln p_X\left(\tilde{x}\right)\right\}. \tag{21}$$

With the assumption of uniform prior on the statistics of the image, one obtain the ML-estimate:

$$\hat{x} = \arg\max_{\tilde{x}\in\Re^N} \left\{p_W\left(y \mid \tilde{x}\right)\right\}. \tag{22}$$

To solve problems (21) and (22) we will refer to the considered above non-stationary Gaussain and stationary Generalized Gaussian models. One can classify the possible image denoising methods into ML (no prior on image) and MAP (with image prior) estimates. An overview of the denoising methods depending on the image and watermark statistics is shown in Figure 11.

6.1.1 ML solution of image denoising problem

The ML-estimate (22) has a closed form solution for several cases when the watermark has either a Gaussian, a Laplacian, or a mixture of Gaussian and Laplacian distributions. If the watermark has a Gaussian distribution the ML-estimate is given by the local mean of y: $\hat{x} = localmean\left(y\right).$

On the other hand, if the watermark can be modeled by a Laplacian distribution the solution of the ML-estimate is given by the local median: $\hat{x} = localmedian\left(y\right)$. In the theory of robust statistics, the mixture model of the Gaussian and Laplacian distributions (so called $\epsilon-$contaminated model) is used. The closed solution in this case is the local trimmed mean filter that uses order statistics such as the median filter but produces the trimmed version of the mean centered about the median point. The size of the window used for the mean computation is determined by the percentage of the impulse outliers given by parameter ϵ hence the name "$\epsilon-$contaminated". If the wa-

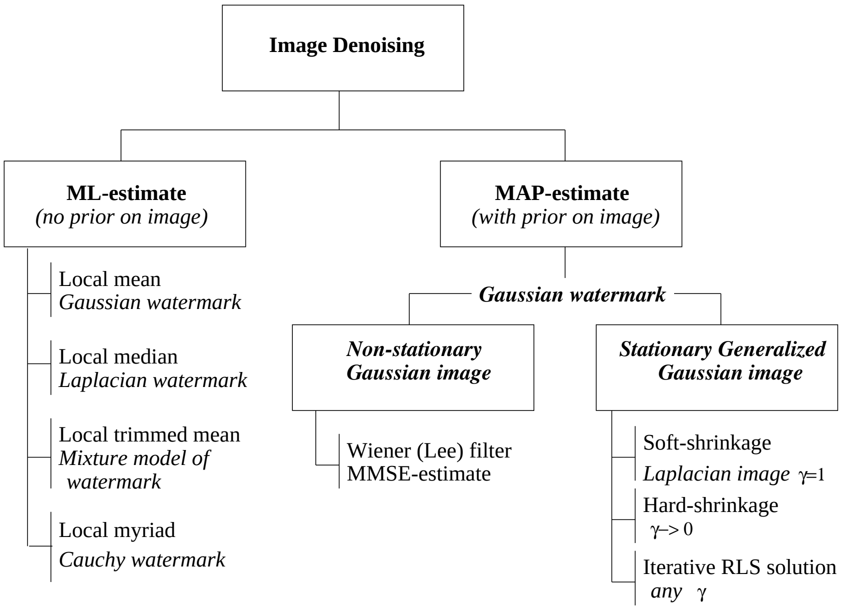


Fig. 11. Classification of image denoising methods.

termark distribution is Gaussian, the ML-estimate results in the myriad filter [31].

In practice, a sliding square window is used in which either the local mean or median is computed. However, in the case of natural images one can compute more accurate estimates of the local mean or median by considering only pixels in a cross-shaped neighborhood. This is due to the fact that natural images feature a higher correlation in the horizontal and vertical directions.

6.1.2 MAP solution of image denoising problem

Assuming $w \sim i.i.d.\ N(0, R_W)$, $R_W = \sigma_w^2 I$ the MAP problem (21) is reduced to [68]

$$\widehat{x} = \arg\min_{\tilde{x}\in\Re^N} \left\{ \frac{1}{2\sigma_w^2} \left\| y - \tilde{x} \right\|^2 + \rho\left(res\right) \right\} \tag{23}$$

where $\rho(res) = [\eta(\gamma)\, |res|]^{\gamma}$, $res = \frac{x-\overline{x}}{\sigma_x}$, $\|.\|$ denotes the matrix norm, and $\rho(res)$ is the energy function for the sGG model.

In practice, iterative algorithms are often used to solve the above problem. Examples of such algorithms are the stochastic [23] and deterministic annealing (mean-field annealing) [17], graduated nonconvexity [2], ARTUR algorithm [49] or its generalization [3]. Of course, it is preferable to obtain the closed form solution for the analysis of the obtained estimate. To generalize the iterative approaches to the minimization of the non-convex function (23) we propose to reformulate it as a *reweighted least squares (RLS)* problem. Then equation (23) is reduced to the following minimization problem [68]:

$$x^{k+1} = argmin_{\tilde{x} \in \Re^N} \{ \frac{1}{2\sigma_n^2} \|y - \tilde{x}^k\|^2 + \phi^{k+1} \|r^k\|^2 \}, \tag{24}$$

where

$$\phi^{k+1} = \frac{1}{r^k} \rho'(r^k), \tag{25}$$

$$r^k = \frac{x^k - \bar{x}^k}{\sigma_x^k}, \tag{26}$$

$$\rho'(r) = \gamma [\eta(\gamma)]^\gamma \frac{r}{\|r\|^{2-\gamma}}, \tag{27}$$

and k is the number of iterations. The main idea of the RLS consists in the replacement of the non-convex function of image priors on the quadratic for a fixed weighting function ϕ. Since the distortion prior or noise distribution functional is also quadratic one can obtain a convergent minimization. Therefore, for specified number of iterations function ϕ is fixed in (24). The obtained solution x^k is then substituted in (26) and a new step of optimization is performed with the new update of the weighting function ϕ (27). This is repeated utill the global minimum is found. We can also obtain closed form solutions for several important image priors, as described below.

First, consider the model: $x \sim N\left(\overline{x}_j, \sigma_{x_j}^2\right)$, $w \sim N\left(0, \sigma_w^2 I\right)$. The solution to this problem is the well known adaptive Wiener or Lee filter:

$$\hat{x} = \overline{y} + \frac{\sigma_x^2}{\sigma_x^2 + \sigma_w^2} \left(y - \overline{y}\right). \tag{28}$$

Second, we assume that $x \sim sGG\left(\overline{x}, 1, \sigma_x^2 I\right)$, i.e. Laplacian, and $w \sim N\left(0, \sigma_w^2 I\right)$. The solution to this problem is soft-shrinkage [68] which is well known in the wavelet domain [20]

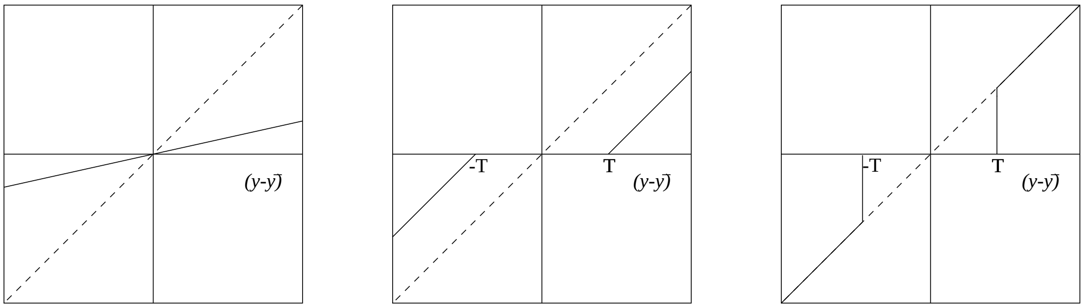
$$\hat{x} = \overline{y} + \max\left(0, |y - \overline{y}| - T\right) sign\left(y - \overline{y}\right) \tag{29}$$

where $T = \sqrt{2} \frac{\sigma_w^2}{\sigma_x}$. It was shown recently [44] that the hard-shrinkage denoiser can be determined under the same priors in the limiting case $\gamma \to 0$:

$$\hat{x} = \overline{y} + \psi\left(|y - \overline{y}| > T\right)\left(y - \overline{y}\right) \tag{30}$$

where $\psi\left(.\right)$ denotes a thresholding function that keeps the input if it is larger than T and otherwise sets it to zero. The main idea of all the above denoisers

(28-30) is to decompose the image into a low frequency part $\overline{y}$ and a high frequency part $(y - \overline{y})$. Each part is then treated separately. The scaling part of the Wiener solution is depicted in figure 12a, and shrinkage functions for soft- and hard-thresholds are shown in figures 12b and 12c, respectively. Relatively small values of $(y - \overline{y})$ represent the flat regions (the same statement is true for wavelet coefficients), while the high amplitude coefficients belong to the edges and textures. Therefore, denoising is mostly due to the "suppression" of noise in the flat regions where the resulting amplitude of the filtered image is either decreased by a local factor $\frac{\sigma_x^2}{\sigma_x^2+\sigma_w^2}$ as in the Wiener filter or just simply equalized to zero as in the case of shrinkage methods. The obvious conclusion is that the shrinkage methods behave in a more aggressive way with respect to the removal of watermark coefficients from the flat image regions, in comparison to the Wiener filter which only decreases their strength. Therefore, it is possible either to remove the watermark in the flat regions completely or to decrease considerably its energy. It is also necessary to note that since the watermark is removed or its strength is decreased the MMSE is decreased while the perceptual quality is enhanced after attack.



(a) (b) (c)

Fig. 12. Scaling/shrinkage functions of the image denoising algorithms

6.2 Lossy wavelet compression attack and its relationship to denoising

The modern wavelet lossy compression algorithms exploit both intra- and interscale redundancy of real images. A well-known example of intrascale model based methods that exploits the zero-correlation across the subbands is the embedded zerotree wavelet (EZW) algorithm proposed in [63] and its extended implementation in SPHIT [61]. The example of interscale coder is EQ-coder proposed by Lopresto et al [41] that utilizes the above stochastic image models for quantization scheme design.

The aim of this section is to show the connection between image denoising and lossy compression with respect to the watermark removal problem. The idea of using lossy compression for denoising has been proposed originally in [62,46]. There are two main recent points of view on this subject based on the work [40,8]. The first approach [40] refers to a theory of complexity regularization and the second one is the generalization of the shrinking principle to the case of quantized data [8]. In our formulation, lossy compression aims to remove the watermark in (3) giving the closest estimate to the cover image in compressed form.

The complexity regularization has the following formulation. Given a measurement $y \in Y$ one should estimate $x \in X$ for a given probabilistic transition model $p(y \mid x)$ that coincides in formulation with (21). However, there is a constraint that the estimate $\widehat{x}$ should be in a discrete set $\Gamma = \{\tilde{x}_j, 1 \leq j \leq J\}$. A codeword is assigned to each of the candidates $\tilde{x}_j \in \Gamma$, so that the estimate is in compressed form. The estimation can be performed based within the MAP paradigm (21)

$$\widehat{x} = \arg\max_{\tilde{x}\in\Gamma} \left\{ \ln p_W(y \mid \tilde{x}) + \ln p^{\Gamma}(\tilde{x}) \right\} \tag{31}$$

where $p^{\Gamma}(\tilde{x})$ is some prior over Γ. It is not the same as the MAP estimate (21) due to the new constraint $\tilde{x} \in \Gamma$ instead of $\tilde{x} \in \Gamma$. This form can be rewritten as [40]:

$$\widehat{x} = \arg\min_{\tilde{x}\in\Gamma} \left\{ -\ln p_W(y \mid \tilde{x}) + \ell(\tilde{x}) \right\} \tag{32}$$

where $\ell(\tilde{x})$ is a length of codeword assigned to $\tilde{x}$ that represents the *complexity* of $\tilde{x}$ in nants (1 nant= $1/ln2$ bits). For our linear model (3) and with the assumption about Gaussian distribution of the watermark the complexity regularization can be rewritten as

$$\begin{aligned}
\widehat{x} &= \arg\min_{\tilde{x}\in\Gamma} \left\{ \frac{1}{2(\ln 2)\sigma_w^2} \|y - \tilde{x}\|^2 + \ell(\tilde{x}) \right\} \\
&= \arg\min_{\tilde{x}\in\Gamma} \left\{ \|y - \tilde{x}\|^2 + 2(\ln 2)\sigma_w^2 \ell(\tilde{x}) \right\}
\end{aligned} \tag{33}$$

Therefore, the obtained estimate is a compressed version of the stego image that satisfies the trade-off $\lambda = 2(\ln 2)\sigma_w^2$ between rate $R = \ell(x)$ and distortion $D = \|y - x\|^2$ [40]. The conclusion is that if the watermark variance can be estimated from the stego image or is bounded by visibility constraints, it is possible to compress the image with automatically chosen regularization parameters using some advanced coders that will satisfy the $R(D)$ condition. Practically this means that the data from the stego domain Y will be mapped into the domain X based on some quantization transform. The main results from complexity regularization applied to watermarking attacks result in the best watermark removal with respect to the given measure of distortion.

A different approach [8] states that denoising is mainly due to the zero-zone in quantization and that the full precision of the thresholded coefficients is of secondary importance. The thresholding rule is derived based on the same sGG image model that was used in our modeling. The denoised coefficients are then quantized outside of the zero-zone based on Risannen's Minimum Description Length (MDL) principle. Therefore, the approximation of the shrinking function is performed as in Figure 13. The uniform threshold quantizer (UTQ) was

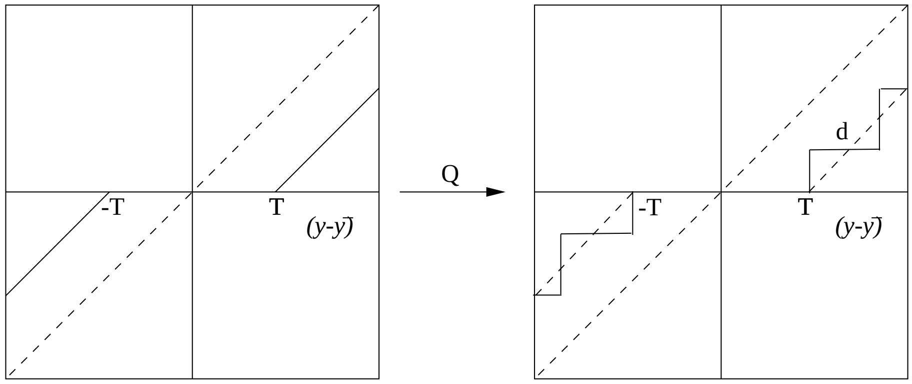


Fig. 13. Approximation of the soft-shrinkage function by quantization with zero-zone.

proposed since it achieves nearly the performance of the entropy-constrained quantizer while being simpler in design [8]. The main difference between the above two approaches is that Liu and Moulin [40] recommended the use of any reasonable coder for denoising while Change *et al* [8] by contrast suggest that the main effectiveness of using compression for denoising is due to the zero-zone in the compression schemes. Our own experiments show that the compression algorithms in the wavelet domain with UTQ show good performance in denoising applications, as exposed further in the next section.

7 Attacks based on estimate of the watermark

Having estimated the watermark the attacker can apply a set of attacks to the stego data. The set of attacks is quite wide and includes the remodulation, the copy attack and the synchronization removal attacks.

7.1 Remodulation attack

The remodulation attack aims at the modification of the watermark using a modulation opposite to one used for the watermark embedding to create the problems for the watermark decoder. The remodulation attack has several different variations depending on the used watermarking decoder. The typical cases include correlation based detection that originates from the ML-detection concept, M-ary modulation and decoding based on error correction codes (ECC). However, independently from the used modulation the watermark is embedded in the image according to a spread spectrum modulation and replicated over the image. The matched filter of the decoder, that performs diversity reception projects the extracted watermark on the key-dependent diversity functions (12) used for the embedding. This results in a coherent reception of the watermark with the increased SNR.

Because the estimated watermark is correlated with the actual watermark, the estimated watermark can be exploited to trick a watermark detector. As shown in Figure (14), the estimated watermark is amplified by a gain factor and then subtracted from the watermarked data. There are four basic variations of the remodulation attack. First, when the gain factor equals 1, the attack yields the MMSE/MAP estimate of the original and reduces to the denoising attack. Secondly, the estimated watermark is amplified by a gain factor larger than 1 and then subtracted from the watermarked data assuming that the watermark was embedded uniformly without perceptual masking [64].

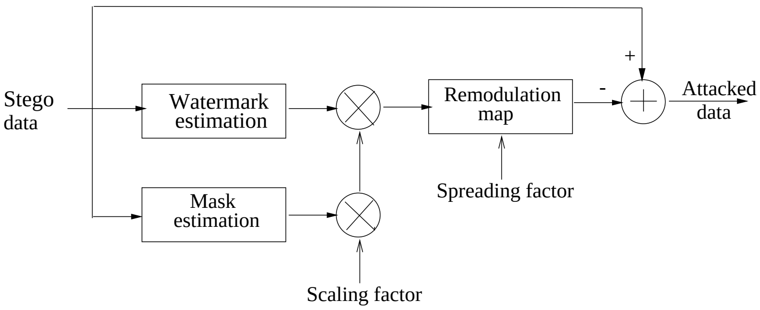


Fig. 14. Perceptual remodulation attack.

By increasing the gain factor, the attack reduces the correlation between the attacked data and the actual watermark; the attack can even drive the correlation to zero so that the detector incorrectly decides that the watermark is not present in the attacked data. Decreasing of the matched filter output will also reduce the SNR that has considerable impact on the final performance of the decoder both for the M-ary modulation and the ECC. Thirdly, more realistic assumption includes the multiplication of the subtracted watermark by the perceptual mask to reduce visual distortions by increasing the gamma. Fourth, the attacker can subtract not only the weighted estimated watermark, but also create outliers to obtain a non-Gaussian noise distribution. It is necessary to note that the linear correlation detector, that is mostly used by the watermarking community, is optimal only for Gaussian noise and is derived based on the ML-detection criterion. Moreover, exploiting features of the HVS the attacker can efficiently embed quite large amount of outliers in the area of the edges and textures without considerable visual distortions. We refer to this attack as perceptual remodulation [69].

The attacker can even combine denoising and perceptual remodulation in one framework to make the attack more efficient. The denoising will remove noise practically from all flat areas reducing output SNR while perceptual remodulation will change the noise statistics leading to non-optimal matched filter performance. This attack has very high efficiency against many watermarking technologies, as will be considered below.

In the case of the QIM, if the detection process is publicly known, the attacker can perturb the public signal in such a way that the attacked signal exactly lies on the decision boundary between different quantizer points. After such an attack, the decoder can only randomly guess whether the received signal sample was originally quantized by quantizer assigned to the codeword 1 or to codeword 2. Thus, the watermark information is completely lost [22].

7.2 Watermark copy attack

The other type of watermark estimation-based attacks is the copy attack [37]. The main idea of this attack is to copy a watermark from one image to another image without knowledge of the key used for the watermark embedding to create ambiguity with respect to the real ownership of data (Figure 15). The difference with the remodulation attacks consists in the final goal which in the previous case is to destroy the watermark and in the second to create the protocol ambiguity.

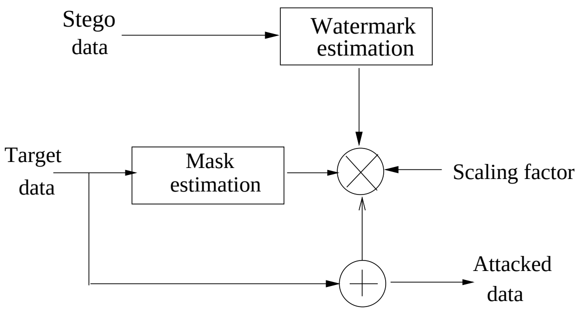


Fig. 15. Copy attack.

The attack consists of two basic stages, i.e. watermark prediction and addition to another image with the adaptation of the predicted watermark features to the target images. As the watermark prediction scheme one can use either the above considered ML or MAP estimates considering the stego image as noisy image with the additive noise to be the watermark.

In the next stage the predicted watermark is adapted to the target image to keep it imperceptible while maximizing the energy. There are many practical ways to adapt the watermark to the target image based on the methods exploiting the contrast sensitivity and texture masking phenomena of the HVS.

To model texture masking we use the NVF based on the stationary Generalized Gaussian model [68]. The NVF characterizes the local texture of the image and varies between 0 and 1, where it takes 1 for flat areas and 0 for highly textured regions. In addition, it is also proposed to take into account [37] the contrast sensitivity to combine it with the NVF. The contrast sensitivity is described by the Weber-Fechner law, which states that the detection threshold of noise is approximately proportional to the local luminance. The final weight is then given by:

$$M = ((1 - NVF)\,\alpha + NVF\,(1 - \alpha))\,Lum, \tag{34}$$

where α describes the relation between the watermark strength in the textured areas and flat areas, and Lum is the local luminance. If we set $\alpha = 1$, the watermark will be concentrated in the texture areas, while taking $\alpha = 0$, the watermark will be mainly embedded in the flat areas.

The fake watermarked image is then generated by scaling the weighted function M, multiplying it by the sign of the predicted watermark, and then adding the result to the target image:

$$y' = t + \beta M sign\,(\widehat{w}) \tag{35}$$

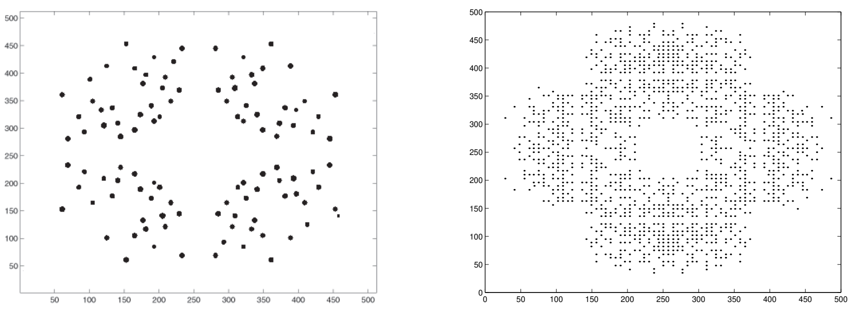
where t is the target image and β is the overall watermark strength.

It is necessary to note that the copy attack in the published version is mainly applicable to the linear additive schemes. In this case the embedding of the watermark in the target image reduces to the simple additive operation. In the case of the QIM, the attacker should first estimate the quantizer used for the embedding and then apply the same quantization table to the target image. The fact that QIM modulation is not image-dependent simplifies the estimation of the used quantizer even without knowledge of the key when the watermark was replicated over the image.

7.3 Synchronization removal attack

Synchronization is a key issue of digital watermarking and the synchronization attacks can be considered as a separate important class of attacks. We concentrate on two main methods of watermark synchronization based on the template in the magnitude image spectrum and the ACF of periodically extended watermark. The main idea of our approach is to detect synchronization mechanisms by analysis of the magnitude spectrum of the predicted watermark $|\Im(\widehat{w})|$. The main assumption is that with state of the art technologies, synchronization is largely based on generating periodic structures. Two possibilities exist. The first consists of inserting peaks in the DFT which is the so called "template" approach used recently by Pereira and Pun to recover from affine transformations [51]. The second approach consists of directly embedding the watermark periodically as done by Kutter [35] in the coordinate domain using ACF and more recently by Voloshynovskiy *et al* [67] in the wavelet domain using magnitude spectrum. In both cases peaks are generated in the DFT which can be exploited by an attacker.

It is obvious that the template peaks and the peaks due to the replicated watermark will be easily detected since the spectrum of periodically repeated watermark has a discrete structure with the period inversely proportion to the period of watermark in the coordinate domain.



(a) template (b) periodic watermark

Fig. 16. DFT peaks associated with a template based scheme (a) and DFT peaks associated with a perceptually embedded periodic watermark(b).

Figure 16 contains an example of detected synchronization in the magnitude image spectrum used in the template approach (a) and the periodic watermark (b).

As an example of this idea, we extracted peaks from the watermarked images based on the template principle used by Digimarc, as shown in Figure 16a. Once the peaks have been detected, the next step of desynchronization is to interpolate the spectrum of the stego or attacked image in the locations of spatial frequencies determined by a local peak detector. We use a simple neighborhood interpolation scheme. As a consequence, any affine geometrical transforms will destroy the watermark synchronization and leave the watermark undetectable. The generalized block diagram of synchronization removal attack is shown in Figure 17.

An example of synchronization removal attack is shown in Figure 18 for the

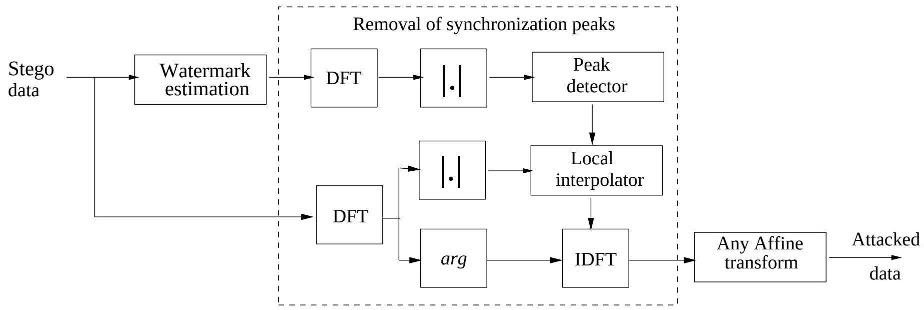


Fig. 17. Synchronization removal attack.

Digimark algorithm integrated in Photo Shop. The software has failed to detect the watermark after this attack followed by a small rotation (about 1 degree). It is necessary to note that the PSNR was increased by about 4 dB after attack. The attack introduces more seveare distortions in the case of the periodical watermarks. This is explained by the larger errors due to the interpolation trying to remove all replicated peaks in the magnitude spectrum of the image. In this case the PSNR is decreased. In the case of the periodical watermark proposed in [67] the watermark was succesfully detected after this attack in 50 images of different size. Therefore, this attack is very efficient mostly against template based synchronization.

(a) (b)

Fig. 18. a) The watermarked image by Digimark watermarking algorithm (PSNR=34.9 dB) b) the attacked image after peaks removal (PSNR=39.2 dB).

8 Possible countermeasures against estimation-based attacks

To resist against estimation-based attacks the information hider should play a game with the attacker consisting of making the watermark unpredictable. There are two basic concepts that focuses on the above issue: Power-Spectrum Condition (PSC) [64] and Noise Visibility Function [68].

Power-Spectrum Condition (PSC): An idealized theoretical approach for analyzing estimation-based attacks treats the cover data and watermark as independent, zero-mean, stationary Gaussian random processes. The watermarked data is the sum of these two processes. Since the original data is given, its power spectrum is assumed fixed, but the watermark power spectrum can be adapted to the cover data. The PSC answers the question of how to match the power spectrum of the watermark with the power spectrum of the cover data to resist against estimation-based attack [64].

The basic idea of PSC relies on the fact that the estimation-based attack uses the MMSE estimator to estimate the watermark. Practically, it is implemented as the substraction of the output of the Wiener filer from the stego data. The mean-squared error (MSE) between the watermark and the estimated watermark is used to measure how well a watermark resists to the estimation. It is shown that the MSE is maximized if and only if the watermark power spectrum is directly proportional to the power spectrum of the original signal. This requirement is called the power spectrum condition (PSC) [64]. A watermark whose power spectrum satisfies the PSC is more resistant against estimation.

An important consideration of the estimation-based attacks is the attacked-data distortion. If distortion is measured by the mean-squared difference between the attacked data and the unwatermarked, original data, then the PSC has another important consequence: To drive the correlation to zero, the attack must also make the distortion as large as the power of the original data, so the attacked data is unlikely to be useful [64].

Noise Visibility Function (NVF): The PSC is attractive because it can be proven rigorously and has a convenient mathematical form. However, its idealized assumptions are not always fulfilled by real-world data. For image watermarking, image denoising provides a natural way to develop practical estimation-based attacks. The watermarked image is treated as a noisy version of the original/host image, and the watermark represents noise that should be eliminated. Thus, the estimated watermark is the same as the estimated noise.

The assumptions behind the PSC can be relaxed in at least two ways according to the consideration provided above. One can treat the original image as a non-stationary Gaussian process or as a stationary Generalized Gaussian process. The noise/watermark can be treated as one of these processes. We will assume that it is still a stationary Gaussian process. In the first case, the denoising method results in the adaptive Wiener filter, while in the second case it reduces to the hard-thresholding and soft-shrinkage as the particular cases.

Both denoising methods produce a texture masking function (TMF), which is derived from the image statistics and is therefore image-dependent. The TMF takes on values in [0,1]; the value of the TMF gauges the sensitivity of the human visual system (HVS) to noise in different image regions. Larger values of the TMF indicate greater noise sensitivity. The HVS is very sensitive to noise in flat image regions, where the TMF approaches unity, and denoising smoothes the image. In contrast, the HVS is very insensitive to noise in highly textured regions or near edges, where the TMF approaches zero and the image is left almost unaltered.

To embed a watermark that resists such estimation, the watermark embedding should use the inverted function that is known as a noise visibility function, defined by NVF = 1 - TMF [68]. The NVF values near unity indicate texture or edge regions where the watermark should be amplified, while NVF values near zero (flat regions) indicate flat regions where the watermark should be attenuated. In this way, the watermark is embedded to resist estimation-based attacks.

The NVF for the practically important case of the non-stationary Gaussian image model has the following form:

$$NVF(i,j) = \frac{1}{1+\sigma_x^2(i,j)}, \tag{36}$$

where $\sigma_x^2(i,j)$ denotes the local variance of the image in a window centered on the pixel with coordinates (i,j), $1 \leq i,j \leq M$. Therefore, the NVF is inversely proportional to the local image energy defined by the local variance. In order to estimate the local image variance the *maximum likelihood (ML)* estimate can be used.

For the stationary GG model is defined as [68]:

$$NVF(i,j) = \frac{w(i,j)}{w(i,j)+\sigma_x^2}, \tag{37}$$

where $w(i,j) = \gamma[\eta(\gamma)]^{\gamma}\frac{1}{||r(i,j)||^{2-\gamma}}$ and $r(i,j) = \frac{x(i,j)-\bar{x}(i,j)}{\sigma_x}$.

The particularities of this model are determined by the choice of two parameters of the model, e.g. the shape parameter γ and the global image variance σ_x^2.

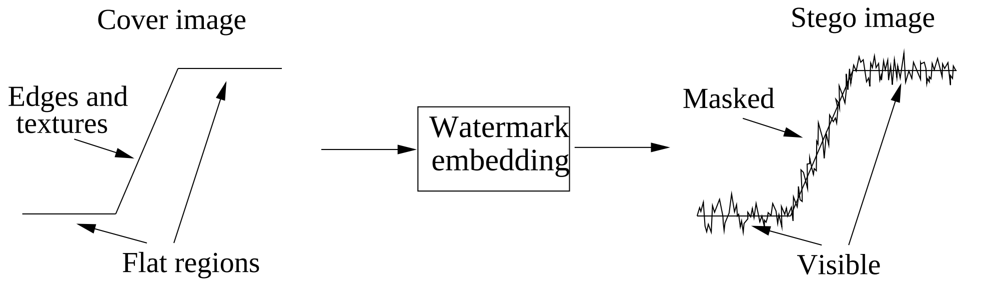
9 Generalized attack: denoising/compression watermark removal followed by perceptual remodulation

Since the watermark being once embedded propagates through "the communication channel", the attacker has more advantages in winning hiding-attacking game. Therefore, the attacker can design more powerful attacks by integrating knowledge of the used watermarking receiver and perceptually constrained measure of distortions. Moreover, in many cases the attacker is not restricted in time to design and to perform the attack oppositely to the hider which could be requested to perform embedding on-line in many applications.

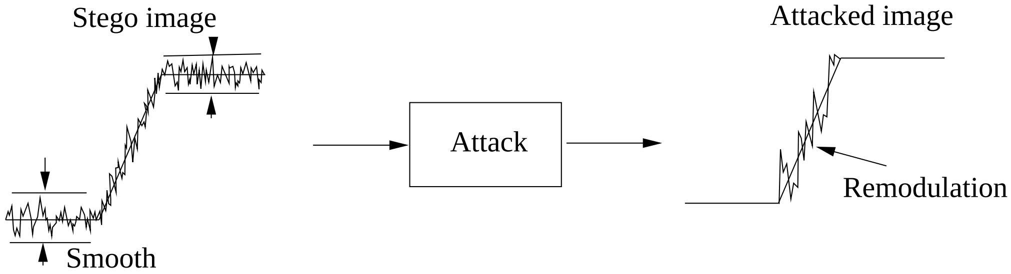
In particular, a simple attack can be designed exploiting the features of the HVS and the watermark predictability. Consider the strategy of the data hider and the attacker in more details. The data hider aims at the designing the optimal embedding strategy maximizing capacity of the channel based on two principles. First, the data hider is going to exploit the masking properties of the HVS and embed as strong as possible watermark in the perceptually invisible image regions. For instance, the texture masking property of the HVS states that the noise/watermark is more visible in the flat regions rather in the areas of the edges and textures (Figure 19a). Second, the data hider will try to make the watermark unpredictable using the PSC or the NVF to escape the possibility of the estimation-based attack application. These two conditions can be easily resolved using the NVF or the PSC as it was considered above.

The attacker will be motivated to reduce the capacity or the rate of reliable communication also exploiting the features of the HVS and the possibility to remove or kill watermark based on different models of the image. Practically it is known that the watermark can be easily predicted and therefore removed from the flat image areas rather than from edges and textures. This fact is based on a very simple model of the image for the flat regions that can be approximated by a local mean. The stochastic model for the edges and textures is non-stationary and much more complicated to be used for the accurate image prediction. Moreover, smoothing distortions in the flat areas are perceptually invisible for the HVS that is commonly exploited in the image compression and denoising applications. Oppositely, the smoothing distortions in the edges and textures are quite visually unpleasant. Therefore, the attacker will tryto utilize the advantages of denoising and to remove the watermark from the flat areas without visual distortions and even enhancing PSNR. Oppositely, the attacker will use the remodulation with the increased strength in the edges and texture arias, which are masked by the HVS like in the case with the data hider, reducing the performance of the matched filter in the watermark receiver. At the same time, the attacker can use the NVF to automatically determine the flat regions, edges and textures. This attack is schematically shown in Figure 19b.

In the case of additive linear watermarking which uses binary phase shift keying (BPSK) modulation, the attacker has to prevent the estimation of the correponding watermark sign and to change it on opposite. This however has to be done for a fraction of the pixels, otherwise one would get a flipping of the watermark which could be easily retrieved. It is thus necessary to change



a) The data hider embedding strategy.



b) The attacker strategy: denoising followed by perceptual remodulation.

Fig. 19. The data hider strategy exploting the texture masking function of the HVS (a) and the attacker strategy using denoising and perceptual remodulation.

the signs randomly (or periodically if some information about the ACF is available) so as to create the least favorable situation for the decoder. There are two different ways to attain this goal.

The first possibility is to estimate the watermark and then to perform remodulation in such a way that the projection of the watermark on the space p in (12) will be on average a zero-mean vector. The particular cases of this generalized attack were studied in [38,25] assuming that the watermark is extracted from the stego image with some strength factor. These attacks have several drawbacks in the case of content-adaptive watermarking, where the strength of the watermark is differs as a function of image regions. In these cases the assumption that the watermark as well as the image are zero-mean, wide-sense stationary Gaussian processes is satisfied neither for the content adaptive watermark nor for the images. As a consequence, the extraction of the watermark is applied with the same strength for flat regions and for edges and textures. Therefore, the watermark could just be inverted and non-visibility is not guaranteed here. This indicates that watermark remodulation should be content adaptive.

The second possibility consists of creating outliers with a sign opposite to the local estimated watermark sign, taking into account visibility constraints [69]. Considering the prior reduction of sampling space in the flat regions due to denoising/compression, this will lead to an unsatisfactory solution when the

CLT assumption is made. The resulted distribution of errors due to outliers will no longer be strictly Gaussian. In this case, the decoder designed for the AWGN will not be optimal and the general performance of the watermarking system will be decreased. Additionally, if the attacker can discover some periodicity in the watermark structure, this could be effectively used for remodulation to reach the above goal. Since, the behavior of the correlator and sign correlator detectors that are mostly used in watermarking decoders is well studied in [32] we will not concentrate on this point here. We will rather present some practical aspects of remodulation.

One method consists of changing the amplitude relationship among the pixels in a given neighborhood set. In the most general case, one has to solve a local optimization problem of watermark sign change under constraint of minimal visible distortions for every pixel in the set like in the case of perceptual remodulation. Based on practically driven motivations one can assume that only some pixels in a neighborhood set should be changed during the optimization, according to some causal image model, or even considering the value of the central pixel only. This will certainly constrain the level of variability but has the benefit of leading to very simple closed form solutions.

Assume one can have the estimate of the watermark sign based on the predictor (11) as

$$s = sign\left(y - \overline{y}\right). \tag{38}$$

The idea is to remodulate the watermark by a sign opposite to s, according to a perceptual mask that will assign stronger weights for the textures and edges and smaller ones for the flat regions (if the Wiener filter is used for the denoising/compression attack). We have used here the texture masking property of the HVS for this perceptual remodulation based on the NVF) [68]. Other reasonable models could be used here as well. In the case of NVF the resulting attacked image can be written as:

$$y = \widehat{x} + \left[(1 - NVF) \cdot S_e + NVF \cdot S_f\right] \cdot (-s) \cdot p' \tag{39}$$

where $\widehat{x}$ is the denoised image, and S_e and S_f are the strengths of the embedded watermark for edges and textures and for flat regions, respectively, $p' \in \{0, 1\}$ is a spreading function for non-periodical watermark with probability of appearance "0" equal to ω, and "1" - $(1 - \omega)$. The block diagram of the described attack is shown in Figure 20. The performance of the attack is demonstrated below.

To investigate the effectiveness of the proposed attack we performed tests for three different watermarking embedding approaches, using 15 gray scale images of size 256x256. Here, we only report the results for Girl image. The tested watermarking algorithms were: method A - a coordinate domain al-

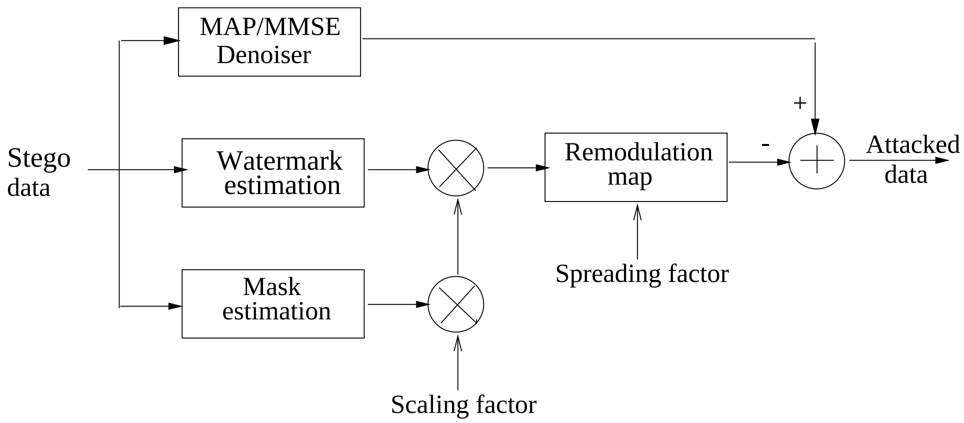


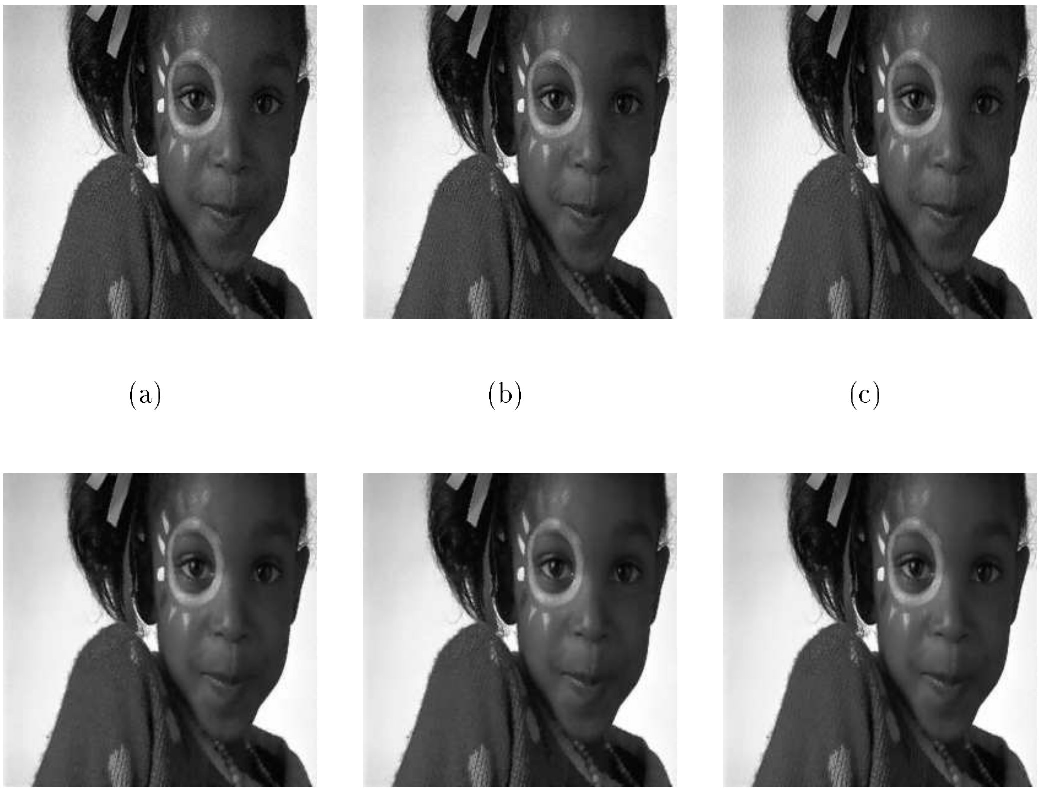
Fig. 20. The block diagram of the generalized attack that consists of denoising followed by perceptual remodulation.

gorithm with ECC encoding and texture masking, and a message length of 64 bits; method B - a coordinate domain algorithm with M-ary modulation (M=2) and luminance masking, 64 bits; method C - a DCT domain method with ECC encoding and Just Noticeable Difference (JND) masking, 48 bits. The corresponding stego images are shown in Figure 21(a-c).

We applied the proposed attack (39) with the Wiener filter as the denoiser and fixed parameters $S_e = 4\sigma_w$, $S_f = 1.05$ and $\omega = 0.3$ for all methods. The variance of the watermark was estimated only from the flat image regions based on the NVF computed according to non-stationary Gaussian image model. The resulted images are shown in Figure 21(d-f). The peak signal-to-noise ratio (PSNR) was chosen to estimate the image quality that is presented in the captions of the corresponding images. To better imagine the strategy of watermark modification according to the proposed attack the watermarks before and after attack are shown in Figure 22 and the corresponding histograms of watermark are shown in Figure 23. In all cases the watermarking softwares indicated that the watermark was not found in the image. In addition, the bit error rate was in the range 60-80% indicating the inability of the algorithms to recover the embedded message. It is also necessary to note that although the PSNR after attack is slightly reduced due to the outliers in the edges and textures, the visual quality of the image is not degraded since the remodulation is performed only in the perceptually invisibly areas.

10 Denoising with Random Bending

Since one the most effective attack to date is the Stirmark random bending attack, we propose an improved version of this attack. Instead of immediately applying random geometric distortions, we choose to first apply soft thresholding. This is done in order to effectively suppress the watermark in flat areas



(a) (b) (c)

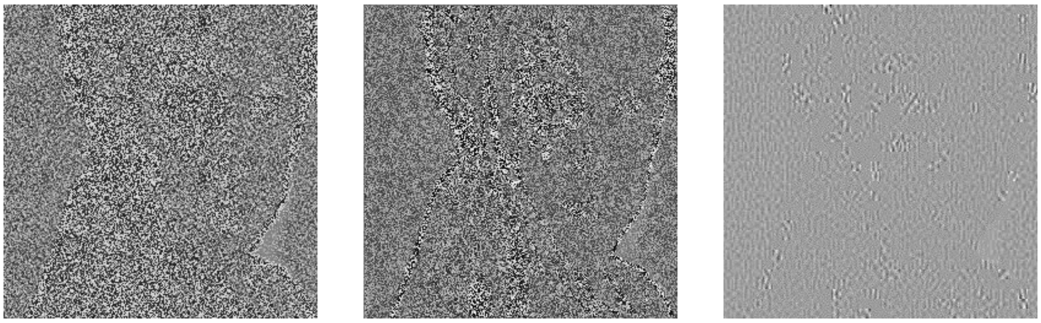
(d) (e) (f)

Fig. 21. The stego Girl image watermarked using method A (a) PSNR=35.34dB, B (b) PSNR=36.77dB and C (c) PSNR=40.77dB, and corresponding attacked images (d-f) with PSNRs equal to 35.52dB, 35.85dB and 38.51dB.

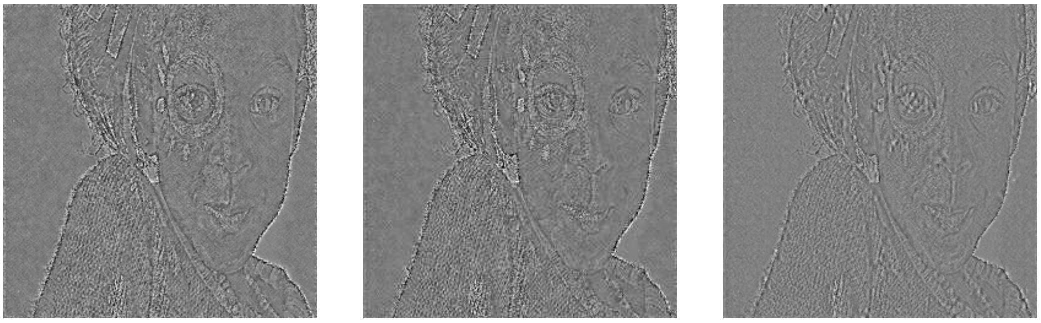
with little or no impact on visual quality. The Stirmark random bending attack is then applied to the denoised image.

11 Perceptual quality estimation

In order to reduce the visibility effects of the insertion of a watermark, algorithms can take advantage of the HVS characteristics by inserting watermarks in the less sensitive regions of the images, such as the textured regions. A good image quality metric should take into account the HVS characteristics to provide accurate measurements and therefore objectively state whether or not a given watermark is visible. Unfortunately the widely used PSNR metric does not take into account such characteristics and it cannot be used as a reference metric for measuring image quality. In fact the PSNR metric does not take into account image properties. The attractiveness of PSNR in appli-

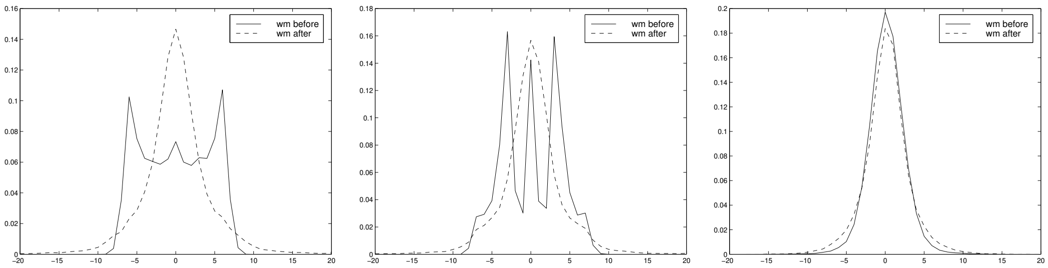


(a) (b) (c)



(d) (e) (f)

Fig. 22. The original watermarks of method A (a), B (b) and C (c), and corresponding resulting watermarks after attack (d-f).



(a) (b) (c)

Fig. 23. The histograms of the watermarks before (solid line) and after attack (dashed line).

cations such as image restoration and segmentation arises from the fact that it is directly related to the squared error. Since typically algorithms attempt to minimize square error, the PSNR accurately measures to what extent this goal was attained. However, in watermarking applications, the goal is to produce a watermark which is as robust as possible while still being invisible. Within this context, the PSNR is inadequate as will be shown.

In what follows we propose two image quality metric, based on a weighted PSNR and on the Watson model, but adapted for to watermarking applications.

11.1 The weighted PSNR

The classical PSNR quality metric is given by:

$$PSNR = 10 log_{10} \frac{max(x)^2}{||x' - x||^2} \tag{40}$$

where x' is the image under test and x is the original image.

In the above equation the PSNR penalizes the visibility of noise (watermark) in all regions of the image in the same way. However, due to phenomena of contrast masking the visibility of noise in flat regions is higher than that in textures and edges.

Therefore, a simple approach to adapt the classical PSNR for watermarking applications consists in the introduction of different weights for the perceptually different regions oppositely to the PSNR where all regions are treated with the same weight. Originally this idea was presented by Netravali and Haskell [47] with application to image compression. In application to watermarking quality evaluation it was reported in [69] using the NVF as a weighting matrix:

$$wPSNR = 10 log_{10} \frac{max(x)^2}{||x' - x||^2_{NVF}} = 10 log_{10} \frac{max(x)^2}{||NVF(x' - x)||^2}. \tag{41}$$

11.2 The Watson model

The central aim of the Watson metric [70] is to weight the errors for each DCT coefficient in each block by its corresponding sensitivity threshold which is a function of the contrast sensitivity, luminance masking and contrast masking.

For a given DCT component (i, j) we have the visibility threshold given by:

$$log_{10} t_{ij} = log_{10} \frac{T_{min}}{r_{ij}} + S(log_{10} f_{ij} - log_{10} f_{min})^2 \tag{42}$$

$$\text{with:} r_{ij} = r + (1 - r) cos^2 \theta_{ij} \tag{43}$$

where the parameters T_{min}, S, and f_{min} are functions of the total luminance of the display L, i.e. background luminance on the screen plus luminance contributed by the image. The parameters T_{min}, S, and f_{min} are given by:

$$T_{min} = \begin{cases} \frac{L}{S_0} & ifL > L_T \\ \frac{L}{S_0}(\frac{L_T}{L})^{1-a_t} & ifL \leq L_T \end{cases} \tag{44}$$

$$S = \begin{cases} k_0 & ifL > L_k \\ k_0(\frac{L}{L_k})^{a_k} & ifL \leq L_k \end{cases} \tag{45}$$

$$f_{min} = \begin{cases} f_0 & ifL > L_f \\ f_0(\frac{L}{L_f})^{a_f} & ifL \leq L_f \end{cases} \tag{46}$$

where $L_T = 13.45cd/m^2, S_0 = 94.7, a_t = 0.649, L_k = 300cd/m^2, k_0 = 3.125, a_k = 0.0706, L_f = 300cd/m^2, f_0 = 6.78cycles/deg, r = 0.7$ and $a_f = 0.182$. Also,

$$f_{ij} = \frac{1}{16}\sqrt{(i/W_x)^2 + (j/W_y)^2} \tag{47}$$

where W_x and W_y are the horizontal and vertical size of a pixel in degrees of visual angle. The angular parameter is given by:

$$\theta_{ij} = arcsin\frac{2f_{i0}f_{0j}}{{f_{ij}}^2} \tag{48}$$

The parameters were determined by extensive subjective tests and are now widely adopted.

Luminance: It has now been established that there is an important interaction between luminance and frequency which Watson incorporates in the model by setting

$$t_{ijk} = t_{ij}(\frac{c_{00k}}{\overline{c}_{00}})^a_t \tag{49}$$

where c_{00k} id the DC coefficient of block k, $\overline{c}_{00}$, and a_t determines the degree of masking (set to 0.65 typically).

Texture: Texture masking refers to the fact that the visibility of a pattern is reduced by the presence of another in the image. The masking is strongest when both components are of the same spatial frequency, orientation and location. Watson extends the results of luminance and frequency masking presented above to include texture masking. This is done by setting:

$$m_{ijk} = Max[t_{ijk}, |c_{ikj}|^{w_{ij}} {t_{ijk}}^{1-w_{ij}}] \tag{50}$$

where m_{ijk} is the masked threshold and w_{ij} determines the degree of texture masking. Typically $w_{00} = 0$ and $w_{ij} = 0.7$ for all other coefficients.

The perceptual error in each frequency of each block if given by:

$$d_{ijk} = \frac{e_{ijk}}{m_{ijk}} \tag{51}$$

where e_{ijk} is the quantization error.

Now to obtain a total perceptual error (TPE) independent of the image size, we pool errors over space and frequency by using the formula:

$$TPE = \frac{1}{N^2} \sum_{k} \sum_{i,j} |d_{ijk}| \tag{52}$$

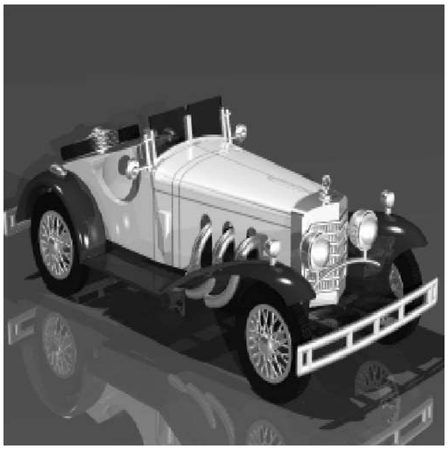
We note that this pooling differs from the Minkowski summation proposed by Watson, however our tests indicate that with respect to the watermarking application better results are obtained.

11.3 Comparison of the Watson metric and the wPSNR with the PSNR

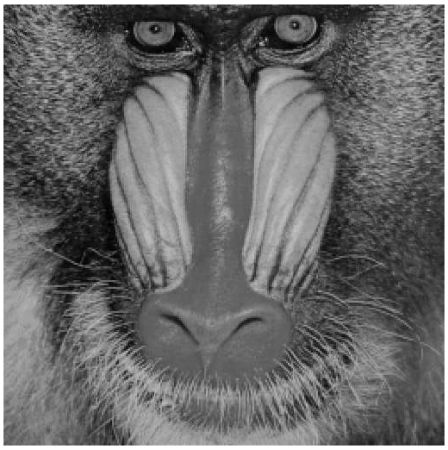
In this section we present some examples that demonstrate the accuracy of the Watson metric in cases where the PSNR is inadequate.

For the first example we add the same additive white noise to the images Benz and mandrill. The PSNR for both images is the same but the visual quality for the mandrill image is much better than for the Benz image as we can see in Figure 24. The Watson metric states that the quality for the mandrill image is much better than for the benz image, which is in accordance with our perception. In order to objectively specify if an image is acceptable or not, we must specify some thresholds beyond which the image is declared unacceptable relative to the proposed objective measure. We determine the following thresholds by performing subjective tests for different types of images and then taking the average values.

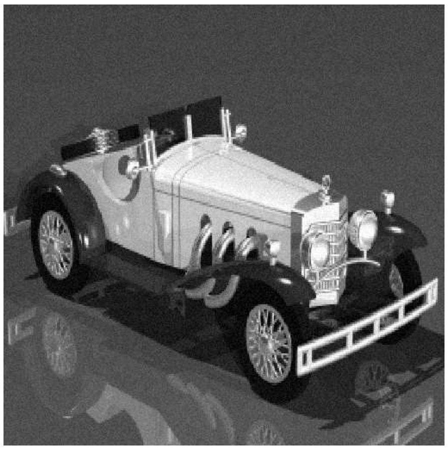
(1) A global perceptual error threshold $GT = 4.1$, so that images with a total perceptual error less than this threshold are considered to be globally of good quality. We find however that in some cases, even though the global threshold is satisfied, the image is too distorted to be of commercial value. This may arise in cases where the watermark has been inserted



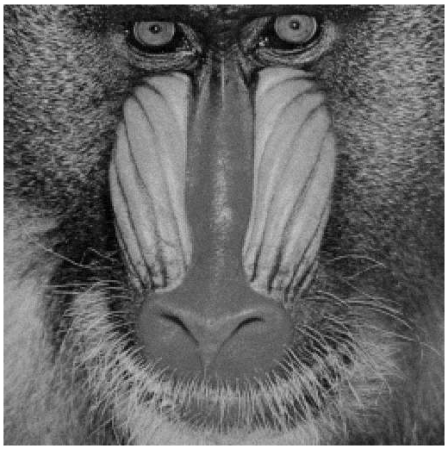
(a) original benz



(b) original mandrill



(c) PSNR=31dB, TPE=9



(d) PSNR=31dB, TPE=4

Fig. 24. Comparison of Benz and Mandrill at constant PSNR=31dB.

too strongly at a few locations. While this may only slightly influence the global criteria of a large image, the watermark will still be visible locally. Consequently we propose also using local measures.

(2) A first local perceptual error threshold $LT1 = 7.6$ for blocks of size 16×16. Blocks with greater total perceptual error than this threshold may be locally visible but not enough to systematically reject the image. The variable NB1 contains the number of these potentially visible blocks. This reflects the fact that the metric is not pefect and that in some cases human judgment is necessary.

(3) A second local perceptual error threshold $LT2 = 30$ for blocks of size 16×16, so that the error in blocks with greater total perceptual error than this threshold are in all cases visible so that the image cannot be accepted. The number of these blocks is reported in the variable NB2 and

Table 1
PSNR and Watson measures for the images Benz and Mandrill

	Same additive white noise			
	PSNR (dB)	TPE	NB1	NB2
benz	31.71	9.07	522	0
mandrill	31.69	4.15	28	0

the image is rejected if NB2 is equal or greater than one.

With respect to the tests, the images were displayed on 24 bit screens from an Ultra Sparc 10. The application used to display the images was XV version 3.10a. It is important to notice that the errors visible to the human eye will depend on the luminance and contrast parameters from the screen and the application used to display the images.

Table (1) reports the PSNR and the Watson measures for the images benz and mandrill from Figure (24). We note that the Watson metric correctly indicates that the errors are much more visible in the Benz image than in the Mandrill image even though the PSNR is the same.

For our second example we consider 3 different watermarked versions of the Barbara image with different image quality but the same PSNR, see Figure 25. The watermarked versions use NVF masking based on non-stationary Gaussian and Generalized Gaussian models of the image and in the third case, no masking is used. Once again the Watson metric provides measurements according to the perceptual reality, thus proving to be more precise than the PSNR. In particular, while the PSNRs are the same, the Watson metric correctly identifies the fact that for the non-adaptive case, the watermark is visible. It is necessary to note, that the wPSNR also correctly classified the subjective quality of the images and shows the same performance as the Watson metric. For our last example we consider the Lena image to which initials of a name have being added and are locally very visible, see Figure 26. The PSNR metric does not say anything about this local degradation, and the total perceptual error is not so useful in this case, but we obtain TPE=0.26, NB1=6, and NB2=2 from the Watson metric. The number of blocks NB2 with greater perceptual error than the second local threshold is 2. Consequently, the image is rejected.

12 Second Generation Benchmarking and Results

Having described various new attacks and having proposed an accurate and objective measure of image quality, we are now in position to define a second

(a)

(b)

(c)

(d)

Fig. 25. a) The original Barbara image b) image watermarked using sGG NVF: PSNR=24.60dB, wPSNR=26.4dB, TPE=7.73, NB1=119, NB2=0 c) image watermarked using ng NVF: PSNR=24.59dB, wPSNR=27.9dB, TPE=7.87, NB1=128, NB2=0, d) watermarked was added to the cover image without masking: PSNR=24.61dB, wPSNR=29.3dB, TPE=9.27, NB1=146, NB2=3

generation benchmark. We note that the benchmark we propose is not intended to replace the benchmark proposed by Kutter and Petitcolas [36], but rather to complement it. While their benchmark heavily weights geometric transformations and contains non-adaptive attacks, the benchmark we propose includes models of the image and watermark in order to produce more effective attacks.

12.1 A New Benchmark

The benchmark consists of six categories of attacks where for each attacked image a 1 is assigned if the watermark is decoded and 0 if not. The categories

(a) (b)

Fig. 26. a) The original Lena image b) initials of a name added

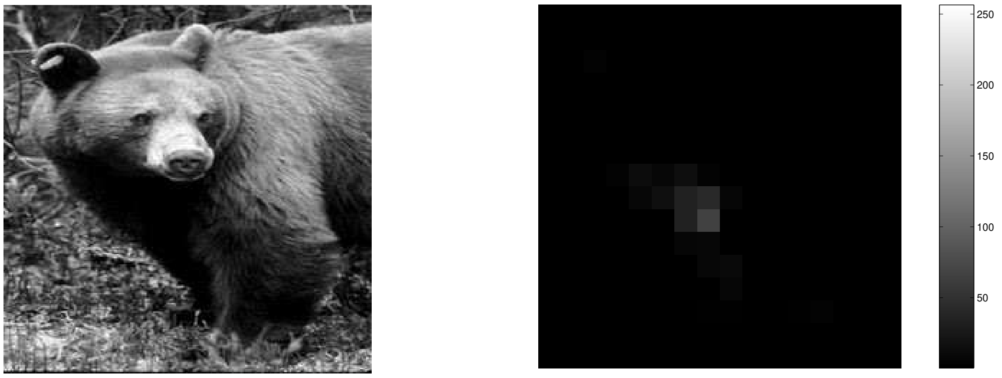
are the following where we note in parentheses the abbreviations we use later for reporting results:

(1) Denoising (DEN): We perform three types of denoising, Wiener filtering, soft thresholding and hard thresholding. We take the average of the three scores.
(2) Denoising followed by perceptual remodulation (DPR) with the parameters of the attack as for the performed experiment in Figure 21.
(3) Denoising followed by Stirmark random bending (DRB).
(4) Copy Attack (CA): We estimate the watermark using Wiener filtering and copy it onto another image. If the watermark is successfully detected in the new image, 0 is assigned otherwise a score of 1 is obtained.
(5) Template removal followed by small rotation (TR).
(6) Wavelet Compression (WC): In this section we compress the image using bitrates [7,6,5,4,3,2,1,0.9,0.8,0.7,0.6,0.5,0.4,0.3,0.2]. We weight the samples between 7 and 1 by 75% while the rest count for 25%. The finer sampling at smaller bitrates is important since most current algorithms survive until a bitrate of 1 or 2 and then start to break down. The finer sampling allows us to better localize at which point the algorithms break down. In some applications such as video, bitrates in the range of 0.2 are frequently encountered. We note that this corresponds roughly to a JPEG quality factor of 10% however the artifacts are much less problematic since the blocking effects do not occur with wavelet compression.

12.2 Results

In this section we report the results relative to the proposed benchmark for two commercial software packages which we denote A and B as well as the algorithm (C) proposed by Pereira in [52]. Algorithm A is a coordinate domain method additive watermark using teture masking. Algorithm B is a DCT domain approach using just noticeable difference masking. Algorithm C is a non-adaptive watermark in the DCT domain which uses texture masking based on NVF.

Table (2) shows Watson measures for 5 images for the watermarks generated by the three approaches. According to this table the error visibility produced by the three watermarking algorithms is locally visible for the bear image to the point that the image is rejected. Figure (27) shows the marked version of the bear image for software C and the total perceptual errors for blocks of size 16×16. We notice that the errors on the marked images are not visible when printed but they are clearly visible on the screen. It is important to note that the errors were not visible under all viewing conditions, but in practice image watermarks must be invisible under all conditions that might be encountered in practice. The Watson metric identified that for all three approaches the watermark was visible in the dark flat areas of the bear image. This is depicted in figure 27. For the rest of the images the Watson metric



(a) (b)

Fig. 27. a) The bear marked image b) total perceptual errors for blocks 16×16

reports a good quality which is in accordance with our observations on the screen.

Table 3 reports the scores of the three algorithms relative to the benchmark. The results were averaged over 5 images. We note that the maximum possible score is 6. The results indicate that the algorithm C based on [52] performs markedly better than the other commercial softwares tested. This results from the fact that algorithm C uses non-linear technique. Such watermarks will be inherently more resistant to the attacks proposed. While it is true that the attacks proposed in the benchmark target linear additive schemes, it is important to note that developing effective denoising approaches for non-additive and non-linear watermarks is much more difficult which suggests in itself that

Table 2
Watson measures for images bear, boat, girl, lena and watch. All images are of size 256×256.

	A			B			C		
	TPE	NB1	NB2	TPE	NB1	NB2	TPE	NB1	NB2
bear	7.10	29	13	2.48	14	4	1.65	13	4
boat	2.61	1	0	0.98	0	0	0.31	0	0
girl	3.44	5	0	1.31	0	0	0.59	0	0
lena	2.63	0	0	1.02	0	0	0.34	0	0
watch	3.69	11	0	1.35	0	0	0.49	0	0

Table 3
Benchmark Results.

	DEN	DPR	DRB	CA	TR	WC	Total
A	0	0	0	0	0	0.79	0.79
B	0	0	0	1	0	0.75	1.75
C	0.93	0.8	0	1	0	0.79	3.52

effective watermarking algorithms should not be based on the linear additive paradigm. We also note that all algorithms fail against the template removal attack and the denoising followed by random bending attacks which indicates that technologies are still not mature relative to the problem of synchronization.

13 Conclusion

In this article we have formalized the problem of attack modelling with emphasis on the linear additive watermarking model. Better understanding of the mechanisms of possible attacks will lead to the development of more efficient and robust watermarking techniques and as such our results present an important step in this direction. Based on our attacks, we have proposed a new benchmarking tool in which we include new attacks which explicitly model the image and watermark. Furthermore we have proposed a new quality metric which provides a much better objective measure of image quality in the context of watermarking.

ACKNOWLEDGMENTS

We thank Frederic Deguillaume, Alexander Herrigel, and Martin Kutter for many fruitful discussions. This work has been financed by the Swiss Priority Program in Information and Communication Structures, by the ESPRIT OMI project JEDI-FIRE, project CERTIMARK and by DCT-Digital Copyright Technologies Switzerland.

References

[1] Unzign watermark removal software, http://altern.org/watermark/. July 1997.

[2] A.Blake and A.Zisserman. MA: The MIT Press, 1987.

[3] A.Dalaney and Y.Bresler. Globally convergent edge-preserving regularization: An application to limited-angle tomography. *IEEE Trans. on Image Processing*, 7(2):204–221, 1998.

[4] A. Alattar. "smart images" using digimarc's watermarking technology. In Ping Wah Wong and Edward J. Delp, editors, *IS&T/SPIE's 12th Annual Symposium, Electronic Imaging 2000: Security and Watermarking of Multimedia Content II*, volume 3971 of *SPIE Proceedings*, San Jose, California USA, 23–28 jan 2000.

[5] M. Barni, F. Bartolini, V. Cappellini, and A. Piva. A DCT-domain system for robust image watermarking. *Signal Processing*, 66:357–372, 1998.

[6] M. Barni, F. Bartolini, A. De Rosa, and A. Piva. Capacity of the watermark-channel: How many bits can be hidden within a digital image ? In *SPIE: Security and Watermarking of Multimedia Contents*, volume 3657, pages 437–448, San Jose, California, January 1999.

[7] M. Barni, F. Bartolini, A. De Rosa, and A. Piva. A new decoder for the optimum recovery of non-additive watermarks. *IEEE Transactions on Image Processing*, submitted 2000.

[8] S. Chang, B. Yu, and M. Vetterli. Spatially adaptive wavelet thresholding with content modeling for image denoising. In *Proc. of 5th IEEE International Conference on Image Processing ICIP98*, Chicago, USA, October 1998.

[9] B. Chen and G. Wornell. Dither modulation: a new approach to digital watermarking and information embedding. In *SPIE: Security and Watermarking of Multimedia Contents*, volume 3657, San Jose, California, January 1999.

[10] Digimark Corporation. http://www.digimark.com/. January 1997.

[11] M. Costa. Writing on dirty paper. *IEEE Transactions on Information Theory*, 29(3):439–441, May 1983.

[12] I. Cox, J. Killian, T. Leighton, and T. Shamoon. Secure spread spectrum watermarking for images, audio and video. In *Proceedings of the IEEE Int. Conf. on Image Processing ICIP-96*, pages 243–246, Lausanne, Switzerland, 1996.

[13] 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.

[14] I. J. Cox and J.-P. M. G. Linnartz. Some general methods for tampering with watermarks. *IEEE Journal on Selected Areas in Communications*, 16(4):587–593, May 1998.

[15] I. J. Cox, M. L. Miller, and A. L. McKellips. Watermarking as communications with side information. *Proceedings of the IEEE*, 87(7):1127–1141, July 1999.

[16] S. Craver, N. Memon, B. L. Yeo, and M. M. Yeung. Can invisible watermark resolve rightful ownerships? In *Fifth Conference on Storage and Retrieval for Image and Video Database*, volume 3022, pages 310–321, San Jose, CA, USA, February 1997.

[17] D.Deiger and F.Girosi. Parallel and deterministic algorithms from mrfs surface reconstruction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(6):401–412, 1984.

[18] F. Deguillaume, G. Csurka, , and T. Pun. Countermeasures for unintentional and intentional video watermarking attacks. In *IS&T/SPIE Electronic Imaging2000*, San Jose, CA, USA, January 2000.

[19] G. Depovere, T. Kalker, and J. P. Linnartz. Improved watermark detection reliability using filtering before correlation. In *IEEE Int. Conference on Image Processing ́98 Proceedings*, Chicago, Illinois, USA, October 1998. Focus Interactive Technology Inc.

[20] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. In *Biometrika*, volume 81, pages 425–455, 1994.

[21] J. Eggers, J. Su, and B. Girod. A blind watermarking scheme based on structured codebooks. In *Secure images and image authentication, IEE Colloquium*, pages 4/1–4/6, London, UK, April 2000.

[22] J. Eggers, J. K. Su, and B. Girod. Robustness of a blind image watermarking schemes. In *Proc. of IEEE International Conference on Image Processing ICIP2000*, Vancouver, Canada, September 2000.

[23] D. Geman and S. Geman. Stochastic relaxation, gibbs distributions and the bayesian restorations of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14(6):367–383, 1984.

[24] F. Hartung and B. Girod. Watermarking of uncompressed and compressed video. *Signal Processing*, 66:283–301, 1998.

[25] F. Hartung, J. K. Su, and B. Girod. Spread spectrum watermarking: Malicious attacks and counterattacks. In *Proc. SPIE Security and Watermarking of Multimedia Contents 99*, San Jose, CA., January 1999.

[26] J. R. Hernández, M. Amado, and F. Pérez-González. Dct-domain watermarking techniques for still images: Detector performance analysis and a new structure. *IEEE Trans. on Image Processing*, 9(1):55–68, January 2000.

[27] J. R. Hernández and F. Pérez-González. Statistical analysis of watermarking schemes for copyright protection of images. *Proceedings of the IEEE*, 87(7):1142–1166, July 1999.

[28] J. R. Hernández, F. Pérez-González, J. M. Rodrígez, and G. Nieto. The impact of channel coding on the performance of spatial watermarking for copyright protection. *in Proc. ICASSP'98*, 5:2973–2976, May 1998.

[29] J. R. Hernández, F. Pérez-González, J. M. Rodrígez, and G. Nieto. Performance analysis of a 2-D-multipulse amplitude modulation scheme for data hiding and watermarking of still images. *IEEE Journal on Selected Areas in Communications*, 16(4):510–523, May 1998.

[30] M. Holliman, N. Memon, and M. Yeung. Watermark estimation through local pixel correlation. In *IS&T/SPIE Electronic Imaging'99, Session: Security and Watermarking of Multimedia Contents*, pages 134–146, San Jose, CA, USA, January 1999.

[31] S. Kalluri and G. R. Arce. Adaptive weighted myriad filter optimization for robust signal processing. In *Proceedings of the 1996 Conference on Information Sciences and Systems*, Prinston, NJ, USA, March 1996.

[32] S. Kassam. *Signal Detection in Non-Gaussian Noise*. Springer Verlag, 1998.

[33] D. Kundur and D. Hatzinakos. Improved robust watermarking through attack characterization. In *Optics Express*, volume 3, pages 485–490, December 1998.

[34] M. Kutter. Watermarking resistent to translation, rotation and scaling. In *Proc. SPIE Int. Symp. on voice, Video, and Data Communication*, November 1998.

[35] M. Kutter. *Digital image watermarking: hiding information in images*. PhD thesis, EPFL, Lausanne, Switzerland, August 1999.

[36] M. Kutter and F. A. P. Petitcolas. A fair benchmark for image watermarking systems. In *Electronic Imaging '99, Security and Watermarking of Multimedia Contents*, volume 3657, pages 219–239, San Jose, CA, USA, January 1999.

[37] M. Kutter, S. Voloshynovskiy, and A. Herrigel. Watermark copy attack. In Ping Wah Wong and Edward J. Delp, editors, *IS&T/SPIE's 12th Annual Symposium, Electronic Imaging 2000: Security and Watermarking of Multimedia Content II*, volume 3971 of *SPIE Proceedings*, San Jose, California USA, 23–28 jan 2000. (Paper EI 3971-35).

[38] G. C. Langelaar, R. L. Lagendijk, and J. Biemond. Removing spatial spread spectrum watermarks by non-linear filtering. In *Proc. Europ. Signal Processing Conf. (EUSIPCO 98)*, Rhodes, Greece, Sept. 1998.

[39] G.C. Langelaar. Watermark removal software, available under a windows95 at http://www-it.et.tudelft.nl/gerhard/waterm.zip. September 1998.

[40] J. Liu and P. Moulin. Complexity-regularized image denoising. In *Proc. of 4th IEEE International Conference on Image Processing ICIP97*, pages 370–373, Santa-Barbara, CA, 1997.

[41] S. LoPresto, K. Ramchandran, and M. Orhard. Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization framework. In *Data Compression Conference 97*, pages 221–230, Snowbird, Utah, USA, 1997.

[42] MediaSec. http://www.mediasec.com/products/download/. March 2000.

[43] P. Moulin. The role of information theory in watermarking and its application to image watermarking. *Signal Processing*, current issue, 2001.

[44] P. Moulin and J. Liu. Analysis of multiresolution image denoising schemes using generalized-gaussian and complexity priors. In *Proc. IEEE Trans. Info. Theory*, volume 45, pages 909–919, April 1999.

[45] P. Moulin and J. O'Sullivan. Information-theoretic analysis of information hiding. *IEEE Information Theory, preprint, submetted: available from http://www.ifp.uiuc.edu/ moulin*, October 1999.

[46] B. Natarajan. Filtering random noise from deterministic signals via data compression. In *Proc. IEEE Trans. Sig. Proc.*, volume 43, pages 2595–2605, nov 1995.

[47] A. Netravali and B. Haskell. *Digital Pictures Representation and Compression.* Plenum Press, New York, 1988.

[48] J. Oruanaidh and T. Pun. Rotation, scale and translation invariant spread spectrum digital image watermarking. *Signal Processing*, 66(3):303–317, 1998.

[49] P.Charbonnier, L.Blanc-Feraud, G.Aubert, and M.Barlaud. Deterministic edge-preserving regularization in computed images. *IEEE Trans. on Image Processing*, 6(2):298–311, 1997.

[50] S. Pereira, J. J. K. Ó Ruanaidh, F. Deguillaume, G. Csurka, and T. Pun. Template based recovery of Fourier-based watermarks using Log-polar and Log-log maps. In *Int. Conference on Multimedia Computing and Systems, Special Session on Multimedia Data Security and Watermarking*, Juin 1999.

[51] S. Pereira and T. Pun. Fast robust template matching for affine resistant watermarks. In *3rd International Information Hiding Workshop*, Dreseden, Germany, September 1999.

[52] S. Pereira, S. Voloshynovskiy, and T. Pun. Effective channel coding for DCT watermarks. In *ICIP 2000*, Vancouver, Canada, submitted.

[53] Shelby Pereira and Thierry Pun. A framework for optimal adaptive dct watermarks using linear programming. In *Tenth European Signal Processing Conference (EUSIPCO'2000)*, Tampere, Finland, sep 5–8 2000.

[54] A. Perrig. A copyright protection environment for digital images. diploma dissertation. *Ecole Polytechnique Federal de Lausanne, Lausanne, Switzerland*, February, 1997.

[55] F. Petitcolas. http://www.cl.cam.ac.uk/ fapp2/watermarking/stirmark/. In *Stirmark3.0(60)*, 1999.

[56] F. A. P. Petitcolas and R. J. Anderson. Attacks on copyright marking systems. In *2nd International Information Hiding Workshop*, pages 219–239, Portland, Oregon, USA, April 1998.

[57] I. Pitas. A method for signature casting on digital images. In *Proceedings of the IEEE Int. Conf. on Image Processing ICIP-96*, pages 215–218, Lausanne, Switzerland, September16-19 1996.

[58] C. I. Podilchuk and W. Zeng. Perceptual watermarking of still images. In *Proc. Electronic Imaging*, volume 3016, San Jose, CA, USA, February 1996.

[59] C. I. Podilchuk and W. Zeng. Image-adaptive watermarking using visual models. *IEEE Journal on Selected Areas in Communications*, 16(4):525–539, May 1998.

[60] K. Ratakonda, R. Dugad, and N. Ahuja. Digital image watermarking: Issues in resolving rightful ownership. In *IEEE Int. Conference on Image Processing ́98 Proceedings*, Chicago, Illinois, USA, October 1998. Focus Interactive Technology Inc.

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

[62] N. Saito. *Simultaneous Noise Supression and Signal Compression using a Library of Orthonormal Bases and the MDL Criterion.* Eds. New York: Academic, 1995.

[63] J. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41:3445–3462, December 1993.

[64] J. Su and B. Girod. Power-spectrun condition for energy-efficient watermarking. In *IEEE ICIP-99*, Kobe,Japan, October 1999.

[65] T. Tirkel, C. Osborne, and R. van Schyndel. Image watermarking - a spread spectrum application. In *Proc. IEEE Int. Symposium on Spread Spectrum Techniques and Applications*, volume 2, pages 785–789, 1996.

[66] S. Voloshynovskiy. Robust image restoration based on concept of m-estimation and parametric model of image spectrum. In *IEEE, IEE, EURASIP 5th International Workshop on Systems, Signals and Image IWSSIP'98*, pages 123–126, Zagreb, Croatia, June 1998.

[67] S. Voloshynovskiy, F. Deguillaume, and T. Pun. Content adaptive watermarking based on a stochastic multiresolution image modeling. In *EUSIPCO 2000*, Tampere, Finland, September 2000.

[68] S. Voloshynovskiy, A. Herrigel, N. Baumgaertner, and T. Pun. A stochastic approach to content adaptive digital image watermarking. In *Third International Workshop on Information Hiding*, Dresden, Germany, September 29 - October 1st 1999.

[69] S. Voloshynovskiy, S. Pereira, A. Herrigel, N. Baumgärtner, and T. Pun. A generalized watermark attack based on stochastic watermark estimation and perceptual remodulation. In Ping Wah Wong and Edward J. Delp, editors, *IS&T/SPIE's 12th Annual Symposium, Electronic Imaging 2000: Security and Watermarking of Multimedia Content II*, volume 3971 of *SPIE Proceedings*, San Jose, California USA, 23–28 January 2000. (Paper EI 3971-34).

[70] A. B. Watson. DCT quantization matrices visually optimized for individual images. In *Proc. SPIE:Human vision, Visual Processing and Digital Display IV*, volume 1913, pages 202–216. SPIE, 1993.