Proceedings Chapter

Generalized watermark attack based on watermark estimation and perceptual remodulation

VOLOSHYNOVSKYY, Svyatoslav, *et al*.

Abstract

Digital image watermarking has become a popular technique for authentication and copyright protection. For verifying the security and robustness of watermarking algorithms, specific attacks have to be applied to test the proposed algorithms. In contrast to the known Stirmark attack, which degrades the quality of the image while destroying the watermark, this paper presents a new approach which is based on the estimation of a watermark and the exploitation of the properties of Human Visual System (HVS). The new attack satisfies two important requirements. First, image quality after the attack as perceived by the HVS is not worse than the quality of the stego image. Secondly, the attack uses all available prior information about the watermark and cover image statistics to perform the best watermark removal or damage. The proposed attack is based on a stochastic formulation of the watermark removal problem considering the embedded watermark as additive noise with some probability distribution. The attack scheme consists of two main stages: watermark estimation based a Maximum a Posteriori (MAP) approach and watermark [...]

Reference

VOLOSHYNOVSKYY, Svyatoslav, *et al*. Generalized watermark attack based on watermark estimation and perceptual remodulation. In: Ping Wah Wong and Edward J. Delp. *IS\&T/SPIE's 12th Annual Symposium, Electronic Imaging 2000: Security and Watermarking of Multimedia Content II*. 2000.

Generalized watermarking attack based on watermark estimation and perceptual remodulation

Sviatoslav Voloshynovskiy[*ab], Shelby Pereira[a], Alexander Herrigel[b],
Nazanin Baumgartner[b], Thierry Pun[*a]
[a]University of Geneva, Department of Computer Science, 24 rue Général-Dufour,
1211 Geneva 4, Switzerland
[b]Digital Copyright Technologies, Stauffacher-Strasse 149, CH-8004, Zurich, Switzerland

ABSTRACT

Digital image watermarking has become a popular technique for authentication and copyright protection. For verifying the security and robustness of watermarking algorithms, specific attacks have to be applied to test them. In contrast to the known Stirmark attack, which degrades the quality of the image while destroying the watermark, this paper presents a new approach which is based on the estimation of a watermark and the exploitation of the properties of Human Visual System (HVS). The new attack satisfies two important requirements. First, image quality after the attack as perceived by the HVS is not worse than the quality of the stego image. Secondly, the attack uses all available prior information about the watermark and cover image statistics to perform the best watermark removal or damage.
The proposed attack is based on a stochastic formulation of the watermark removal problem, considering the embedded watermark as additive noise with some probability distribution. The attack scheme consists of two main stages: a) watermark estimation and partial removal by a filtering based on a Maximum a Posteriori (MAP) approach; b) watermark alteration and hiding through addition of noise to the filtered image, taking into account the statistics of the embedded watermark and exploiting HVS characteristics. Experiments on a number of real world and computer generated images show the high efficiency of the proposed attack against known academic and commercial methods: the watermark is completely destroyed in all tested images without altering the image quality. The approach can be used against watermark embedding schemes that operate either in coordinate domain, or transform domains like Fourier, DCT or wavelet.

Keywords: watermarking, attacks to watermarks, maximum *a posteriori* estimate, perceptual mask.

1. INTRODUCTION

Digital watermarking has emerged as an appropriate tool for author rights protection [1]. A lack of systematic benchmarking of existing methods however creates confusion amongst content providers and watermarking technology suppliers. Existing benchmarking tools like Stirmark or Unzign integrate a number of image processing operations or geometrical transformations aimed at removing watermarks from a stego image. However, the quality of the attacked image is often too degraded to permit further commercial exploitation. Moreover, the design of such 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 do test, and real world attacks. In this context, the aim of the paper is to present a generalized attack for a wide class of linear additive watermarking technologies; this attack could be integrated in future testing tools.

The first attacking methods which tried to take into account the statistics of images and watermarks are [2, 3]. The main idea of these attacks is to perform watermark estimation and then remodulate the watermark by means of subtracting the estimated watermark from the stego image with some constant strength factor. The strength factor is determined based on either the condition of maximum watermark energy removal, or the condition of minimization of cross-correlation coefficient between the attacked image and the watermark. These attacks have 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.

* Correspondance: Email: {svolos, Thierry.Pun}@cui.unige.ch; WWW: http://cuiwww.unige.ch/~vision

The existing additive linear watermarking methods [4,5,6] have several weak points that could be used by an attacker to design an efficient attack. The key-independent prediction of the watermark makes it possible to perform watermark removal or modification even without knowledge of the key used for the watermark embedding. The removal of the watermark can lead to a decrease of the amplitude sampling space, and to an invalidation of the assumptions about channel noise statistics used for the design of the optimal message decoder. Knowledge of the message decoder structure designed under the assumption of additive Gaussian noise makes it possible to perform a remodulation of the watermark and create the least favorable noise distribution for the decoder.

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 Section 2 with respect to a communication formulation of digital image watermarking, which is decomposed into message embedding and extraction processes. Section 3 presents watermark attacks based on the outlined weak points and their corresponding assumptions. In particular, image denoising/compression is considered as the means to reduce watermark redundancy. Perceptual remodulation is proposed to change the statistics of the noise so that they will not anymore match those expected by the optimally designed decoder. Section 4 contains the results of computer modeling performed for three different watermarking methods and section 5 concludes the paper.

2. PROBLEM FORMULATION

Consider the general model of a watermarking system according to a communication formulation. Its block diagram is shown in Figure 1.

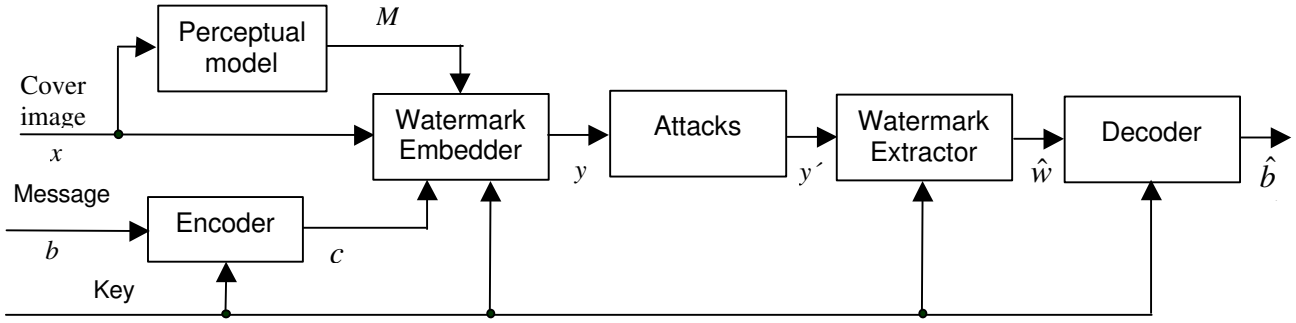


Figure 1. Communication formulation of a watermarking system.

The watermarking 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.

2.1. Message embedding

A message $\mathbf{b}=(b_1,...,b_L)^T$ is to be embedded in the cover image $\mathbf{x}=(x_1,...,x_N)^T$ of size $M_1\times M_2$, where $N=M_1\cdot M_2$. The message $\mathbf{b}$ contains information about the owner and can be used for authentication purposes. To convert the message $\mathbf{b}$ into a form efficient for communication, it is encoded using either error correction codes (ECC) or modulated using some sort of M-ary modulation. 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 $\mathbf{c}=Enc(\mathbf{b},Key)$, $\mathbf{c}=(c_1,...,c_K)^T$ which are mapped from $\{0,1\}$ to $\{-1,1\}$.

A watermark $\mathbf{w}$ is created by some key-dependent function $\mathbf{w}=\varepsilon(\mathbf{c},\mathbf{p},M,Key)$ that ensures the necessary spatial allocation of the watermark based on a key-dependent projection function $\mathbf{p}$, and according to HVS features as expressed by a perceptual mask M in order to improve watermark invisibility. The typical choice for the projection function $\mathbf{p}$ is a set of bidimensional orthogonal functions $P_k=\{(j),\forall p_k(j)\neq 0\}$, $k=1,...,K$ used for every codeword bit $\{c_k\}$ such that the empty set is formed by the intersection $P_K\cap P_l,\forall k\neq l$ [4, 5]. The projection function performs a "spreading" of the data

over the image area that is defined. 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 [7, 8]. A review of various perceptual masks M is given in [9]. The resulting watermark is obtained as the superposition

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

where $j \in \mathbf{Z}^2$.

The watermark embedder performs the insertion of the watermark into the cover image in some key-dependent transform or coordinate domain, yielding the stego image:

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

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

$$\mathbf{y} = h(\mathbf{x}, \mathbf{w}) = \mathbf{x} + \mathbf{w} \tag{3}$$

that is considered in this paper.

The stego image may be subject to different alterations. The alterations could be unintentional or intentional in respect to the removal of the watermark. Here we study possible attacks that can remove the additive robust watermark.

2.2. Message extraction

The message extraction part of the watermarking system consists of the watermark extractor and decoder, as shown in Figure 1; these elements are described below.

2.2.1. Watermark extractor

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

$$\hat{\mathbf{w}} = Extr\left(T_{Key}[\mathbf{y}'], Key\right). \tag{4}$$

In the general case, the extraction should be key-dependent. However, the desire to recover data after affine transformation based on the self-reference principle, and the opportunity to enhance the decoding performance by reducing the variance of the image considered as noise [4, 10], 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 empirical methods are used for watermark estimation, such as the cross-shaped filter [7], or MMSE estimates [5]. In the most general case, the problem of watermark estimation can be solved based on a stochastic framework by using Maximum Likelihood (ML) or MAP estimates [9]. 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:

$$\hat{\mathbf{w}} = \arg\max_{\tilde{\mathbf{w}} \in \Re^N} p_X\left(\mathbf{y}' \mid \tilde{\mathbf{w}}\right) \tag{5}$$

which results either in a local average predictor/estimator in 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.. It is important to note that the cross-shaped predictor is nothing else than:

$$\hat{\mathbf{w}} = C\mathbf{y}' = \mathbf{y}' - \overline{\mathbf{y}'} = \mathbf{y}' - A\mathbf{y}' = (I - A)\mathbf{y}' \tag{6}$$

where C is a cross-shaped high-frequency filter, $\overline{\mathbf{y}}$ is a local mean and A is a low-pass filter with the same support as C.

The MAP estimate is given by:

$$\hat{\mathbf{w}} = \arg\max_{\tilde{\mathbf{w}} \in \Re^N} \{ p_X(\mathbf{y}' \mid \tilde{\mathbf{w}}) \cdot p_W(\tilde{\mathbf{w}}) \} \tag{7}$$

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

$$\hat{\mathbf{w}} = \frac{R_w}{R_w + \hat{R}_x} (\mathbf{y}' - \overline{\mathbf{y}}') \tag{8}$$

where it is assumed $\overline{\mathbf{y}}' \approx \overline{\mathbf{x}}$, and where $\hat{R}_x = \max(0, \hat{R}_y - R_w)$ is the ML estimate of the local variance ($\hat{R}_x = \sigma_x^2 I$).

A key-independent watermark prediction according to (8) presents several problems. One part of problems is connected with the main assumption that the stego image is not significantly altered after attack. This allows to assume that the perceptual mask used for watermark embedding could be estimated from the attacked stego image [4-7]. This assumption does not hold for attacks connected with histogram modification that could have a significant influence on the models based on luminance masking, and lossy JPEG compression attack whose strong blocking artifacts could alter the models based on texture masking.

Another series of problems are tied to general security-robustness issues. Since the watermark could be predicted based on (8) without knowledge of the key using an estimate of the sign $(\mathbf{y} - \overline{\mathbf{y}})$, the following problems appear:

- the image can be modified in such a way that $(\mathbf{y} - \overline{\mathbf{y}})$ produces either "0" or a wrong sign, even with preserving the quality of the attacked image;
- the redundancy in the attacked watermark can be considerably reduced due to the partial watermark removal, this especially in flat image regions;
- special types of distortions could be introduced in the watermark, aiming to create the least favorable conditions for the decoder that will be considered in the next section.

2.2.2. Watermark decoding

Knowledge of the decoder structure is the best way to design a successful attack. In the general case the decoder/demodulator design is based on ML or MAP approaches. Since the appearance of $\mathbf{b}$ 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 detector that performs a despreading of the data in the way of "coherent accumulation" of the sequence $\mathbf{c}$ spread in the watermark $\mathbf{w}$, 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 (Figure 2):

$$\mathbf{r} = \langle \hat{\mathbf{w}}, \mathbf{p} \rangle . \tag{9}$$

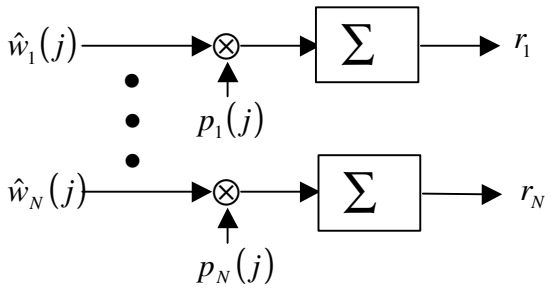


Figure 2. The despreading part of the correlation detector, acting as a dimensionality reduction scheme.

Therefore, given an observation vector $\mathbf{r}$, the optimum decoder that minimizes the conditional probability of error assuming that all codewords $\mathbf{b}$ are equiprobable is given by the ML decoder:

$$\hat{\mathbf{b}} = \arg\max_{\tilde{b}} p\left(\mathbf{r} \mid \tilde{\mathbf{b}}, \mathbf{x}\right). \tag{10}$$

Based on the central limit theorem (CLT) most researchers assume that the observed vector $\mathbf{r}$ can be accurately approximated as the output of an additive Gaussian channel noise [4, 5]

$$\mathbf{r} = A\mathbf{c} + \mathbf{n} \tag{11}$$

where A is a deterministic diagonal matrix composed of a product superposition of the central window value of the prediction filter, and $\mathbf{n}$ is a zero-mean Gaussian random vector. Assuming independence of A and $\mathbf{n}$ the generalized scheme of the decoder for ECC is shown in Figure 3a and for M-ary modulation in Figure 3b.

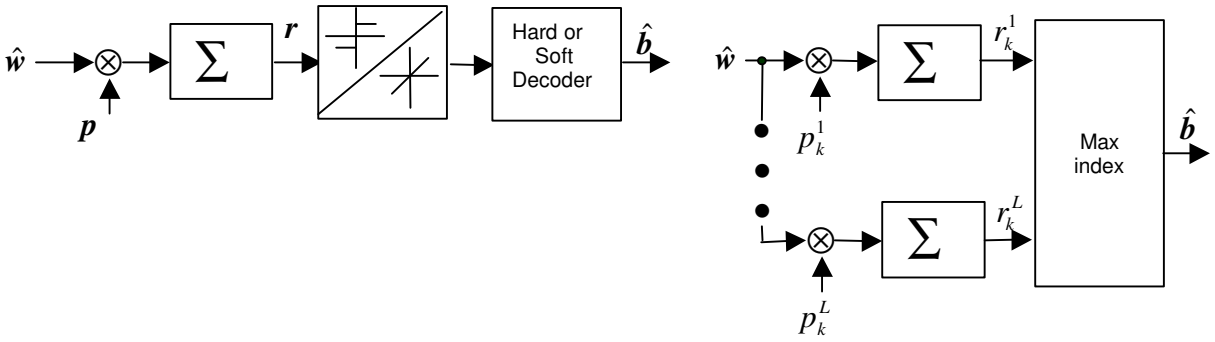


Figure 3. Optimal hard/soft decoder for ECC (a) and for M-ary modulation (b): p_k^i is a set of orthogonal functions used for the allocation of the k-th bit (block) of message, for the L orthogonal sequences chosen according to the key.

However, the above decoders are not optimal for non-Gaussian noise [11]. The knowledge of the decoder structure could therefore be used by the attacker to create the least favorable conditions for message decoding. The weak point of the above approach is the assumption about additive Gaussian noise in (11). This assumption does not hold for the small sample spaces present in the case of small images or due to denoising/compression attacks. In this case the attacker can create non-Gaussian noise in (11) even preserving image quality. Non-Gaussian noise will lead to an increase in the transition probability q in the case of a hard ECC decoder, and to the misinterpretation of noise outliers as the most reliable samples in the codeword in the case of a soft ECC decoder. The creation of non-Gaussian noise will also lead to a non-optimal structure of the matched filter in the case of M-ary modulation. Therefore, the efficiency of both watermarking approaches could be considerably decreased due to this attack.

3. WATERMARK ATTACKS BASED ON THE WEAK POINTS OF LINEAR METHODS

Assuming that the watermark is estimated according to (8), two of the main weaknesses of watermarking algorithms and the corresponding attacks are:

- watermark removal based on denoising/compression that uses the assumption of key-independent watermark extraction aimed at reducing the watermark redundancy;
- perceptual remodulation of the watermark aimed at creating the least favorable statistics for the AWGN decoder designed based on (11).

3.1. Watermark removal based on denoising/compression attack

The watermark can be removed from the stego image using denoising/compression attack. Consider the MAP estimation of the cover image as image denoising according to the additive model (3)

$$\hat{\mathbf{x}} = \arg\max_{\tilde{x}\in\Re^N}\{\ln p_W(\mathbf{y}\mid\tilde{\mathbf{x}}) + \ln p_X(\tilde{\mathbf{x}})\}. \tag{12}$$

To solve this problem it is necessary to develop accurate stochastic models for the cover image $p_X(x)$ and the watermark $p_W(n)$.

3.1.1. Stochastic models of cover image

Stochastic models of cover image applied to content adaptive watermarking were considered in our previous work [9]. 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 like image denoising, restoration and compression, and the existence of tractable closed form solutions of (12) for the particular cases of these models.

3.1.2. 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 [12]. However, for the practically important general case we will constrain our consideration to the stationary Gaussian model to show the connection between the influence of different closed form image denoising methods and of lossy compression on the watermark decoding process. Finally, assuming $\mathbf{w} \sim i.i.d.\quad N(0, R_w)$, $R_w = \sigma_w^2 I$ the MAP problem (12) is reduced to [9]

$$\hat{\mathbf{x}} = \arg\min_{\tilde{\mathbf{x}}\in\Re^N}\left\{\frac{1}{2\sigma_w^2}\|\mathbf{y}-\tilde{\mathbf{x}}\|^2 + \rho(\mathbf{res})\right\} \tag{13}$$

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

3.1.3. MAP solution of image denoising problem

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

$$\hat{\mathbf{x}} = \bar{\mathbf{y}} + \frac{\sigma_x^2}{\sigma_w^2 + \sigma_x^2}(\mathbf{y}-\bar{\mathbf{y}}). \tag{14}$$

Second, we assume that $\mathbf{x} \sim sGG(\bar{\mathbf{x}}, 1, \sigma_x^2 I)$, i.e. Laplacian, and $w \sim N(0, \sigma_w^2 I)$. The solution to this problem is a soft-shrinkage [9] that is well known in the wavelet domain [13]

$$\hat{\mathbf{x}} = \bar{\mathbf{y}} + \max(0, |\mathbf{y}-\bar{\mathbf{y}}| - T)\,sign(\mathbf{y}-\bar{\mathbf{y}}) \tag{15}$$

where $T = \dfrac{\sigma_w^2}{\sigma_x}\sqrt{2}$. It was shown recently [14] that the hard-shrinkage denoiser can be determined under the same priors in the limiting case $\gamma \to 0$:

$$\hat{\mathbf{x}} = \bar{\mathbf{y}} + (\mathbf{y}-\bar{\mathbf{y}})\mathbf{1}(|\mathbf{y}-\bar{\mathbf{y}}| > T) \tag{16}$$

where $\mathbf{1}\{\cdot\}$ 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 (15-16) is to decompose the image into a low frequency part $\bar{\mathbf{y}}$ and a high frequency part

$(\mathbf{y}-\overline{\mathbf{y}})$. Each part is then treated separately. The scaling part of the Wiener solution is depicted in Figure 4a, and shrinkage functions for soft- and hard-thresholds are shown in Figures 4b and c, respectively. Relatively small values of $(\mathbf{y}-\overline{\mathbf{y}})$ 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 $\frac{\sigma_x^2}{\sigma_w^2+\sigma_x^2}$ 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 to remove the watermark in the flat regions, which in turn leads to a decrease of the general watermark redundancy.

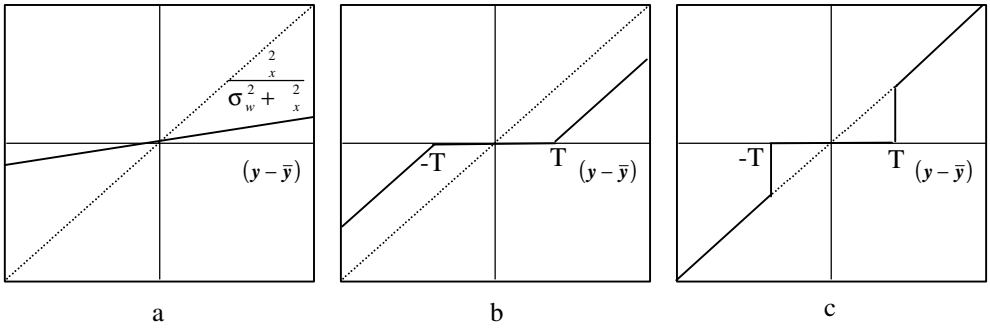


Figure 4. Scaling/shrinking functions of image denoising methods: (a) Wiener/Lee filter; (b) soft-shrinkage; (c) hard-shrinkage.

3.1.4. Watermark removal based on lossy compression

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 [15, 16]. There are two main recent points of view on this subject based on the work [17, 18]. The first approach [17] refers to a theory of complexity regularization and the second one is the generalization of shrinking principle to the case of quantized data [18]. In our formulation, lossy compression aims to remove watermark in (3) giving the closest estimate to the cover image in the compressed form.

The complexity regularization has the next basic problem formulation. Given a measurement $\mathbf{y}\in \mathrm{Y}$ one should estimate $\mathbf{x}\in \mathrm{X}$ for a given probabilistic transition model $p(\mathbf{y}\mid\mathbf{x})$ that coincides in formulation with (12). However, there is a constraint that the estimate $\hat{\mathbf{x}}$ should be in a discrete set $\Gamma=\{\tilde{x}_j, 1\le j\le J\}$. A codeword is assigned to each of the candidates $\tilde{x}_j\in\Gamma$, so that the estimate is in a compressed form. The estimate can be done based on the basic MAP (12)

$$\hat{\mathbf{x}}=\arg\max_{\tilde{x}\in\Gamma}\left\{\ln p_W(\mathbf{y}\mid\tilde{\mathbf{x}})+\ln p^{\Gamma}(\tilde{\mathbf{x}})\right\} \tag{17}$$

where $p^{\Gamma}(\tilde{\mathbf{x}})$ is some prior over Γ. It is not the same as the MAP estimate (12) due to the new constraint $\tilde{\mathbf{x}}\in\Gamma$ instead of $\tilde{\mathbf{x}}\in\Re^N$. This form can be rewritten as [17]:

$$\hat{\mathbf{x}}=\arg\min_{\tilde{x}\in\Gamma}\left\{-\log p_W(\mathbf{y}\mid\tilde{\mathbf{x}})+\ell(\tilde{\mathbf{x}})\right\} \tag{18}$$

where $\ell(\tilde{\mathbf{x}})$ is a length of codeword assigned to $\tilde{\mathbf{x}}$ that represents *the complexity* of $\tilde{\mathbf{x}}$ in bits. For our linear model (3) and with the assumptions done in section 3.1.3 the complexity regularization can be rewritten as

$$\hat{\mathbf{x}}=\arg\min_{\tilde{x}\in\Gamma}\left\{\frac{1}{2(\ln 2)\sigma_W^2}\left\|\mathbf{y}-\tilde{\mathbf{x}}\right\|^2+\ell(\tilde{\mathbf{x}})\right\}=\arg\min_{\tilde{x}\in\Gamma}\left\{\left\|\mathbf{y}-\tilde{\mathbf{x}}\right\|^2+2(\ln 2)\sigma_W^2\,\ell(\tilde{\mathbf{x}})\right\}. \tag{19}$$

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 $\mathbf{R} = \ell(\tilde{\mathbf{x}})$ and distortion $\mathbf{D} = \|\mathbf{y} - \tilde{\mathbf{x}}\|^2$ [17]. 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 $\mathbf{R}(\mathbf{D})$ condition (Figure 5). Practically this means that the data from the stego domain Y will be mapped into the domain X based on some quantization transform (Figure 6). The main results from complexity regularization in application to watermarking attacks are that the best watermark removal according to the given measure of distortion will be performed by a compression scheme which has the operational point determined by a slope of $1/\lambda$.

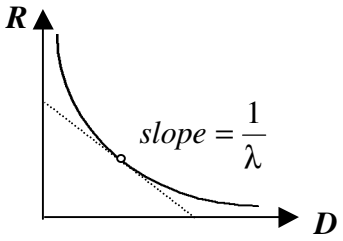


Figure 5. Rate-Distortion curve for some coder with the corresponding operational point.

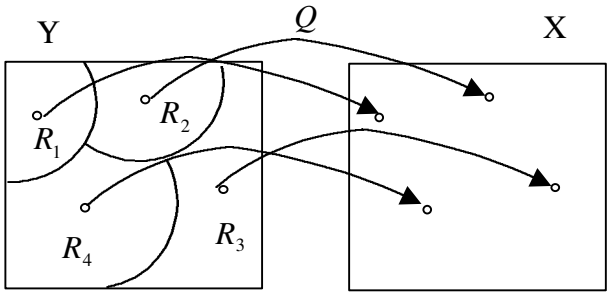


Figure 6. Complexity regularized estimate given in the compressed form based on the quantizer *Q*.

A different approach ([18]) states that the 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 shrinkaging function is performed as in Figure 7.

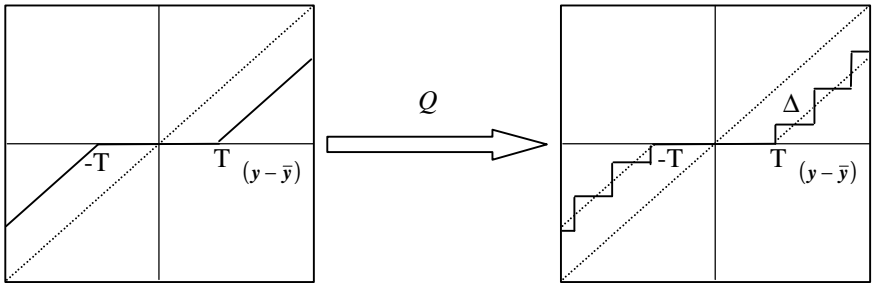


Figure 7. Approximation of the soft-shrinkage function by quantization with zero-zone.

The uniform threshold quantizer (UTQ) was proposed since it achieves nearly the performance of the entropy-constrained quantizer while being simpler in design [18].

The main difference between the above two approaches is that Liu and Moulin [17] recommended to use any reasonable coder for denoising while Change *et al* [18] in 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 the denoising applications, as exposed further in the next section.

3.1.5. Generalized condition for watermark removal based on denoising/compression

Consider the watermark predictor (8). The watermark sign is determined by the difference $(\mathbf{y}-\overline{\mathbf{y}})$, i.e. to have $+1:(\mathbf{y}-\overline{\mathbf{y}})>0$ and $-1:(\mathbf{y}-\overline{\mathbf{y}})<0$. Assume that some neighborhood set of pixels $A\in$ Y was chosen for the estimation of the local mean

$$\overline{y}_i=\frac{\sum_{j\in A_i} y_j a_j}{\sum_{j\in A_i} a_j} \tag{20}$$

where a_j are the weighting coefficients used to model the local image correlation within $a_j\in A$. Assume some denoising/compression method described by a local transform Q was applied to every pixel. After transformation the above condition of sign estimation should hold. However, in the most aggressive case the condition of complete watermark sign removal is

$$Q(y_{c_i})=\frac{\sum_{j\in A_i} Q(y_j) a_j}{\sum_{j\in A_i} a_j} \tag{21}$$

where y_{c_i} is a central pixel for which the sign of watermark should be estimated. Therefore, as a result, the maximum number of uniquely distinguishable watermark signs will be determined by the number of used reconstruction cells in the quantization scheme (Figure 7). This immediately leads to the estimate of the capacity of a given image for a given denoising/compression attack.

Finally, the condition of watermark removal assuming $\{a_j\}=1$ for the chosen neighborhood set is

$$sign(\hat{w})=0, \quad if \;\; y_{c_i}\in R_k, y_j\in R_l \;\; and \;\; k=l \;\; \forall j\in A \tag{22}$$

where R_k are reconstruction cells. The obvious conclusion from (22) is that this condition can be easily satisfied for the flat regions without considerable visual distortions. However, it could cause highly visible changes near edges and in textured regions, leading to a loss in the image values.

3.2. Remodulation of watermark

Another way to prevent the satisfactory estimation of the watermark sign is to change this sign into its opposite. This however has to be done for a fraction only of the pixels, otherwise one would get a flipping of the watermark which could easily be retrieved. It is thus necessary to change 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 reach this goal.

The first way is to first estimate the watermark and then to perform a remodulation in such a way that the projection of the watermark on the space $\mathbf{p}$ (9) will be in average a zero-mean vector. The particular cases of this generalized attack were studied in [2, 3] 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 different for different 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 way consists of creating outliers with a sign opposite to the estimated watermark sign, taking into account visibility constraints. Considering the prior reduction of sampling space in the flat regions due to denoising/compression, this will lead to an unsatisfactory solution of the CLT. The resulted distribution of errors due to outliers will not anymore be

strictly Gaussian. In this case, the decoder designed for AWGN will be not 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 [11] we will not concentrate on this point here. We will rather present some practical aspects of remodulation. A way to do it is to change the amplitude relationship among the pixels in the 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. We call this approach *perceptual remodulation*.

Based on practically driven motivations one can assume that only some pixels in a neighborhood set should be changed during 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 (8) as

$$\mathbf{s} = sign\left(\mathbf{y} - \overline{\mathbf{y}}\right). \tag{23}$$

The idea is to remodulate the watermark by a sign opposite to $\mathbf{s}$, 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 Noise Visibility Function (*NVF*) [9]. Other reasonable models could be used here as well. In the case of *NVF* the resulted attacked image can be written as:

$$\mathbf{y}' = \hat{\mathbf{x}} + \left\lfloor \left(1 - \mathbf{NVF}\right) \cdot S_e + \mathbf{NVF} \cdot S_f \right\rfloor \cdot \left(-\mathbf{s}\right) \cdot \mathbf{p}' \tag{24}$$

where S_e and S_f are the strengths of the embedded watermark for edges and textures and for flat regions, respectively, $\mathbf{p}' \in \{0,1\}$ is a spreading function for non-periodical watermark with probability of appearance "0" equals to ω, and "1" - $\left(1-\omega\right)$. We summarize the proposed ideas for image denoising and perceptual remodulation in the block diagram depicted in Figure 9.

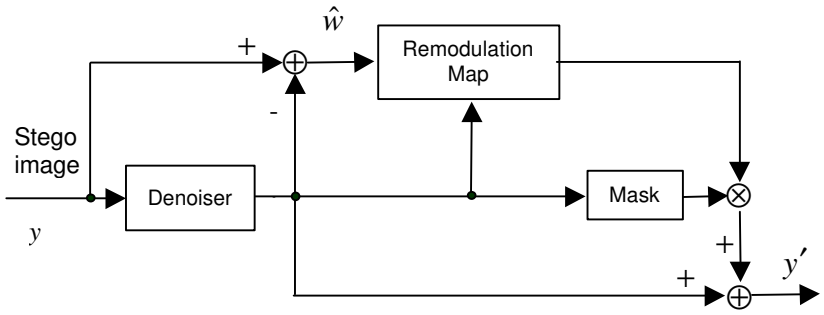


Figure 9. Generalized attack based on key-independent watermark estimation using denoising and perceptual remodulation.

4. RESULTS

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 algorithm 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.

We applied the proposed attack (24) with the Wiener filter as the denoiser and fixed parameters $S_e = 4\sigma_w$, $S_f = 1.05$, $\omega = 0.3$ 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 peak signal-to-noise ratio (PSNR) was chosen to estimate the image quality. PSNR results are shown in Table 1 for the stego, denoised and attacked images for the tested algorithms.

Table 1. PSNR results for several watermarking algorithms before and after the proposed attack

PSNR, dB	A	B	C
Stego image	35.34	36.77	40.77
Denoised image	37.43	38.73	41.36
Attacked image	35.52	35.85	38.51

It is commonly known that the PSNR does not reflect the subjective image quality. Therefore, we used a simple modified measure of image quality based on the weighted PSNR defined as

$$wPSNR = 10\log\frac{\max(\mathbf{x})^2}{\|\mathbf{y}'-\mathbf{x}\|^2_{NVF}} = 10\log\frac{\max(\mathbf{x})^2}{\|(\mathbf{y}'-\mathbf{x})\cdot NVF\|^2}. \quad (25)$$

The main idea to use the *wPSNR* is the fact that the visibility of noise (watermark) on the flat image regions is higher in comparison with the visibility in the regions of textures and edges. Therefore, the relative penalty for this visibility should be decreased for these regions in the measure of image quality. This is efficiently reached based on the NVF which reflects the masking properties of the HVS. The corresponding results for the *wPSNR* are gathered in Table 2.

Table 2. *wPSNR* results for the tested watermarking algorithms before and after the proposed attack

PSNR, dB	A	B	C
Stego image	36.86	37.42	42.42
Denoised image	39.85	40.14	43.37
Attacked image	38.83	38.58	41.32

The results are shown in Figure 10. The stego images are depicted in Figure 10 (a-b) for methods A, B, and C, respectively and the corresponding watermarks are shown in Figure 10 (d-f). The watermarks were obtained as the difference between the stego and cover images. The images after attack are shown in Figure 10 (g-f) and the corresponding watermarks in Figure 10(i-l). The histogram plots of watermarks are presented in Figure 10 (m-o) in solid line before attack and in dot line after attack.

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 algorithm to recover the embedded message.

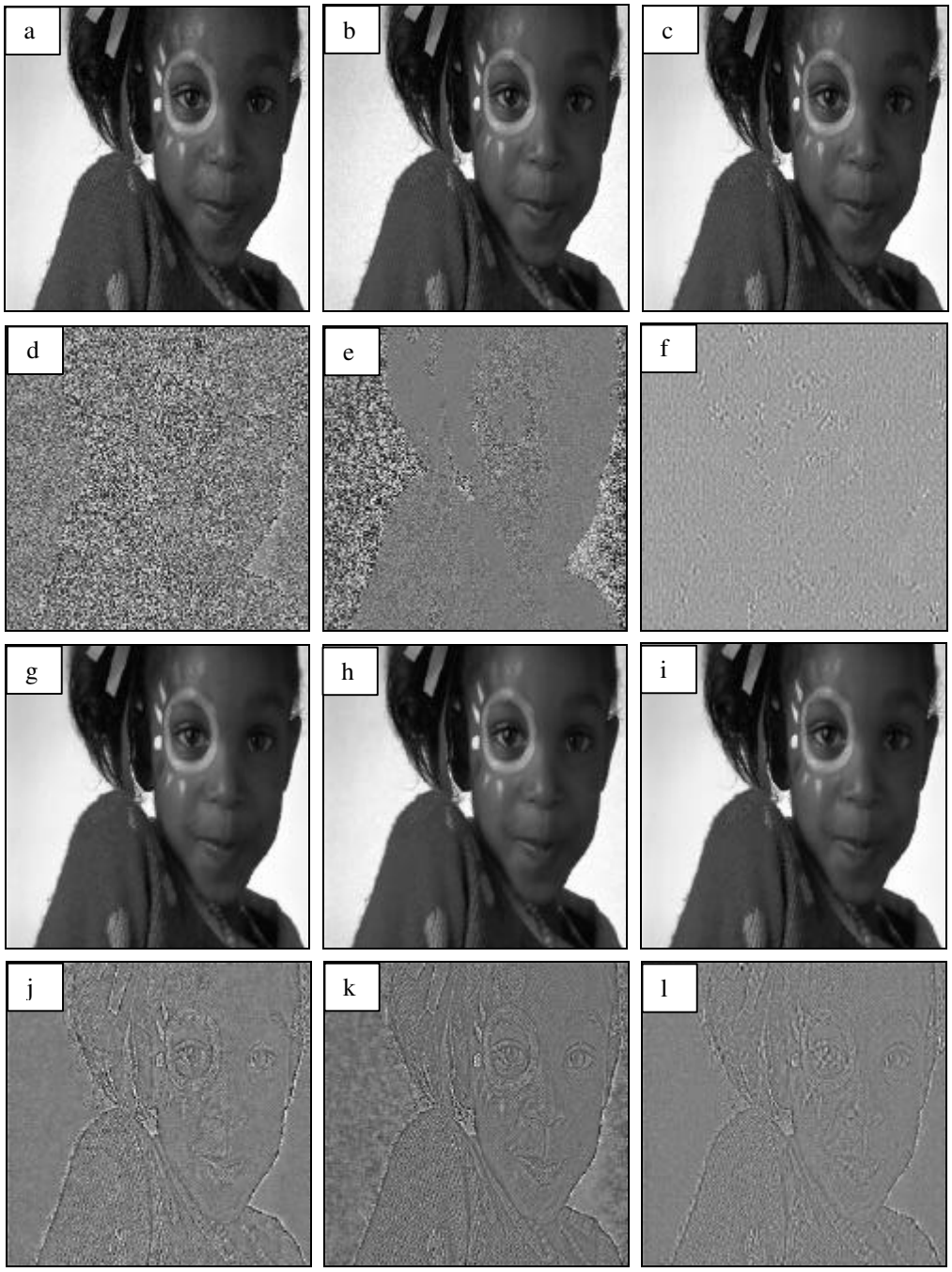
5. CONCLUSIONS

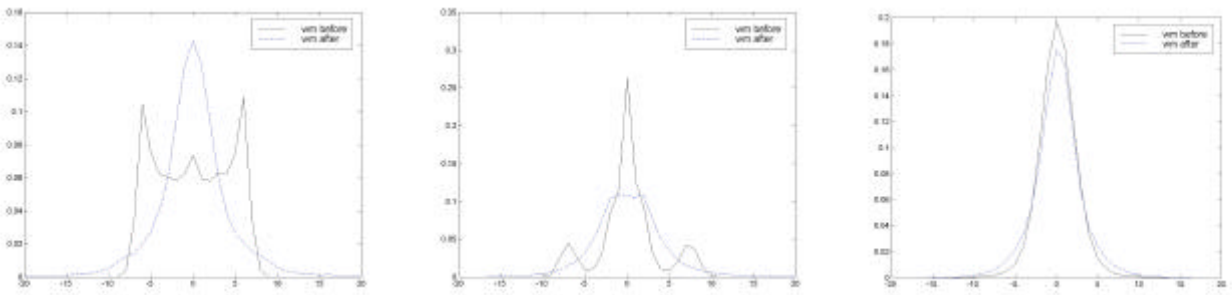
We have presented in this paper a new approach for watermark attacks based on an additive linear watermarking model. In contrast to previous attacks, this approach preserves image quality while simultaneously demonstrating a high efficiency. The approach is based on a denoising/compression paradigm, aiming at reducing the watermark redundancy and remodulating the watermark in order to create the least favorable situation for the decoder. The results confirm the high efficiency of the proposed attack when applied to a number of existing watermarking methods.

6. REFERENCES

1. I. Cox, J. Kilian, T. Leighton, T. Shamoon, "Secure spread spectrum watermarking for multimedia", *IEEE Trans. Image. Processing*, **6**, No 12, pp. 1673-1687, Dec. 1997.
2. G. Langelaar, R. Lagendijk, and J. Biemond, "Removing spatial spread spectrum watermarks by non-linear filtering", *in Proc. EUSIPCO98*, **4**, pp. 2281-2284, 1998.
3. J. Su and B. Girod, "Power spectrum condition for L2-efficient watermarking", *Submitted to IEEE Int Conf. Image Processing ICIP99*, October 1999.
4. M. Kutter, *Digital image watermarking: hiding information in images*, PhD thesis, EPFL, Lausanna, Switzerland, 1999.

5. J. Hernandez, F. Perez-Gonzalez, J. Rodriguez, "Performance analysis of a 2-D-multipulse amplitude modulation scheme for data hiding and watermarking of still images", *Proc. IEEE Journal on Selected Areas in Communications*, **16**, No 4, pp. 510-524, May 1998.
6. J.P. Linnartz, G. Depovere, T. Kalker, "On the design of a watermarking system: cosideration and rationals", 3rd *Workshop on Information Hiding*, Dresden, Germany, Sept. 29-Oct. 1, 1999, to appear in Lecture notes in Computer Science.
7. M. Kutter, "Watermarking resisting to translation, rotation, and scaling", *Proc. of SPIE, Multimedia systems and* applications, **3528**, pp. 523-531, San Jose, USA, November 1998.
8. S. Voloshynovskiy, F. Deguillaume, T. Pun, "Content adaptive watermarking based on a stochastic multiresolution image modeling", submitted EUSIPCO'2000.
9. Voloshynovskiy, A. Herrigel, N. Baumgartner, T. Pun, "A stochastic approach to content adaptive digital image watermarking", 3rd Workshop on Information Hiding, Dresden, Germany, Sept. 29-Oct. 1, 1999, to appear in Lecture notes in Computer Science.
10. J. Hernandez and F. Perez-Gonzalez, "Statistical analysis of watermarking schemes for copyright protection on images", *Proc. IEEE*, **87**, No 7, pp. 1142-1166, July 1999.
11. S. Kassam, *Signal detection in non-Gaussian noise*, Spriner-Verlag, New York, 1998.
12. S. Voloshynovskiy, "Robust Image Restoration Based on Concept of M-Estimation and Parametric Model of Image Spectrum", in *Proc. IEEE, IEE, EURASIP 5th International Workshop on Systems, Signals and Image Processing* ``*IWSSIP'98''*, pp. 123-126, Zagreb, Croatia, June 1998.
13. D. Donoho and I. Johnstone, "Ideal spatial adaptation via wavelet shrinkage", *Biometrika*, **81**, pp. 425-455, 1994.
14. P. Moulin and J. Liu, "Analysis of Multiresolution Image Denoising Schemes Using Generalized Gaussian and Complexity Priors", *Proc. IEEE on Information Theory*, **45**, No3, pp. 909-919, April 1999.
15. N. Saito, "Simultaneous noise suppression and signal compression using a library of orthnormal bases and the MDL criterion", *in Wavelets in Geophysics*, E. Foufoula-Georgiou and P. Kumar, , pp. 299-324, Eds. New York: Academic, 1995.
16. B. Natarajan, "Filtering random noise from deterministic signals via data compression", *IEEE Trans. on Signal Processing*, **43**, No 10, pp. 2595-2605, November 1995.
17. J. Liu and P. Moulin, "Complexity-regularized image denoising", in *Proc. IEEE Int. Conf. Image Processing ICIP97*, **2**, Santa Barbara, CA, pp. 370-373, 1997.
18. S. Chang, B. Yu, and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denosing", *Submitted to IEEE Trans. on Image Processing*, 1998, (http://gabor.eecs.berkley.edu/~grchang/publications.html)





m n o

Figure 10. Results (see text).