Data hiding capacity analysis for real images based on stochastic non-stationary geometrical models

S.Voloshynovskiy, O. Koval, F. Deguillaume and T. Pun

University of Geneva - CUI, 24 rue General Dufour, CH 1211,Geneva 4, Switzerland

ABSTRACT

In this paper we consider the problem of capacity analysis in the framework of information-theoretic model of data hiding. Capacity is determined by the stochastic model of the host image, by the distortion constraints and by the side information about watermarking channel state available at the encoder and at the decoder. We emphasize the importance of proper modeling of image statistics and outline the possible decrease in the expected fundamental capacity limits, if there is a mismatch between the stochastic image model used in the hider/attacker optimization game and the actual model used by the attacker. To obtain a realistic estimation of possible embedding rates we propose a novel stochastic non-stationary image model that is based on geometrical priors. This model outperforms the previously analyzed EQ and spike models in reference application such as denoising. Finally, we demonstrate how the proposed model influences the estimation of capacity for real images. We extend our model to different transform domains that include orthogonal, biorthogonal and overcomplete data representations.

Keywords: watermarking, stochastic image model, capacity, information theory, spike model, edge process model.

1. INTRODUCTION

An important problem of digital data-hiding is the investigation of fundamental theoretical capacity limits, i.e. somehow an analog to Shannon's limit in digital communications. The recent work proposed by Moulin advocates a game-theoretic approach for the evaluation of data-hiding capacity.[1] In the scope of this approach, the data-hiding capacity is considered to be a solution of a max-min game between the data hider attempting at maximizing channel capacity and the attacker aiming at decreasing the capacity. It is also assumed no specific form of encoder/decoder and attacking, but it is rather supposed that both the data hider and the attacker are doing the best to achieve their goals. Therefore, one is looking for the maximum rate of reliable communications, over any possible data-hiding strategy, and any attack that satisfy the specified constraints.

An emerging practical problem is the application of the game-theoretic paradigm for the analysis of data-hiding capacity of real images. The solution to this problem should provide the justification of many existing practical algorithms and allow to fairly evaluate their performance and potential capabilities. An important aspect of this problem is to develop appropriate models for attacking channels, for distortion metrics, and for image statistics. The analysis of these three items has great impact on the solution of the max-min problem. Therefore, it is justified by Moulin that a minimum mean square error (MMSE) estimator of the host signal and a *Gaussian test channel* from rate distortion theory are the worst case attacks for the given constrained attack distortion D_2. A squared-error distortion measure $d(x,y) = (x-y)^2$ is selected for the analysis due to its wide usage in communication theory and nice closed-form results. A Gaussian model of the host image is selected as a source model, since it provides the *upper bound* on capacity for non-Gaussian sources as well. It is also assumed that the maximum admissible distortion for the data hider is D_1 while for the attacker it is constrained by D_2. This abstract scenario of the max-min game is shown in Figure 1.

An important assumption is that both the data-hider and the attacker not only share complete information about their strategies (beside a secret key used by the data-hider for the watermark encryption/encoding/spatial allocation), but also the stochastic model of the source image. In the original analysis of Moulin and Mihcak,[2] it is proposed to use an *expectation − quantization* (EQ)[3] and a *spike model*[4] to evaluate the data-hiding capacity of real images due to their superior performance in the reference applications. Therefore, the final watermark energy allocation and the attack distortion distribution are performed according to the selected models. Although, this approach is

Further author information: (Send correspondence to S. Voloshynovskiy): E-mail: svolos@cui.unige.ch, http://watermarking.unige.ch

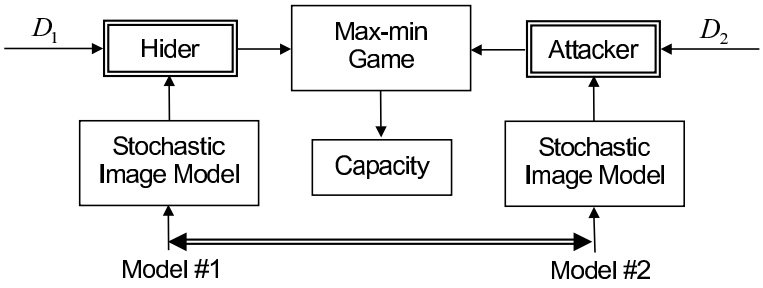


Figure 1. Generalized diagram of game between the data hider and the attacker.

intuitively justified, it could potentially lead to overestimate the actual capacity since the obtained estimate of the rate of reliable communications is highly sensitive to the source model selection. Although the EQ and spike models have a lot of advantages, they do not prevent the possibility that new more powerful source models can appear, and tht the obtained fundamental capacity limits should be determined again. Therefore, the upper limit character of the game approach can be questioned in this case. Moreover, an even worse situation can happen in practice, when the optimal watermark energy allocation is performed assuming the EQ or spike model (denoted as model number 1) in the scope of the game approach, but the actual attacker will apply a new more powerful model for real attacks providing the same attack distortion D_2. This situation is schematically shown in Figure 2. Such a formulation has a lot in common with broadcasting systems and corresponding problems of capacity evaluation.[5]

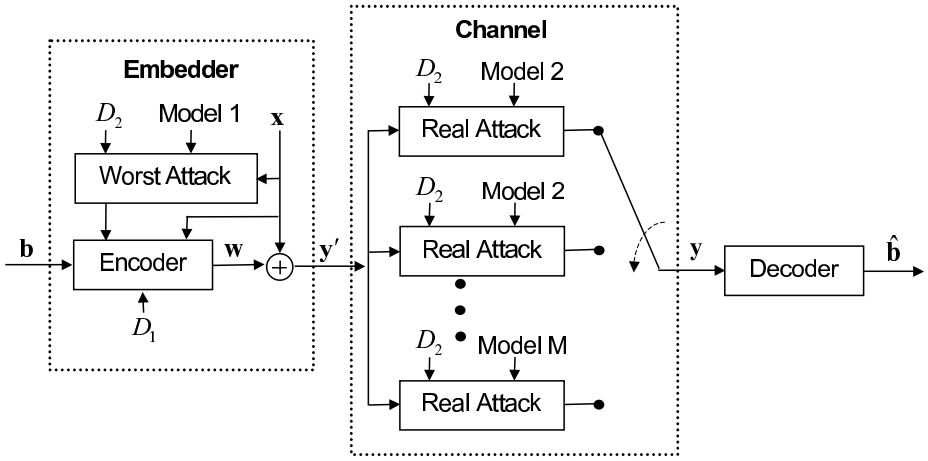


Figure 2. Multichannel game model.

The goal of this paper is to introduce a new class of stochastic image models based on geometrical priors that have superior performance in some reference applications such as denoising as opposed to the EQ and spike models. The proposed model is applied to the game-theoretic set-up and new capacity limit estimates are obtained. Since the proposed model has considerably lower local variances, the obtained practical embedding capacity is also smaller than those obtained for the EQ/spike models. Therefore, we demonstrate that although the information-theoretic game approach is an excellent framework for providing the absolute limits of watermarking capacity, the practical limits are highly sensitive to the model selection and to the transform domain. Essentially, we show that actual capacity might be much lower than predicted by the EQ or spike image models.

Section 2 briefly reviews the main results of Moulin's game approach and new extended data-hiding games are presented in Section 3. A nontrivial problem of image model selection is discussed in Section 4. Section 5 introduces an *edge process* (EP) model and provides the comparison of this model with the EQ and spike models for the reference applications. The actual capacity for the edge process model is presented in Section 6 for several test

images. A new attack based on the edge process model is presented in Section 7. Section 8 summarizes some open issues of game-theoretic approach and concludes the paper.

2. DATA-HIDING GAME-THEORETIC APPROACH

We assume that the message $\mathbf{m}$ is encoded based on a secret key into some watermark $\mathbf{w}$ and embedded into a host data (image) $\mathbf{x}$. The resulting stego data $\mathbf{y}'$ is obtained as:

$$\mathbf{y}' = \mathbf{x} + \mathbf{w}. \tag{1}$$

We include in this generalized model both spread spectrum schemes and host interference cancellation schemes based on pre-coding according to Costa codes. We denote $\mathbf{x}$ to be a two-dimensional sequence representing the luminance of the original image. The *ith* element of $\mathbf{x}$ is denoted as $x[i]$ where $i = (n_1, n_2)$ and $\mathbf{x} \in \mathbb{R}^N$ and $N = M_1 \times M_2$ is the size of the host image. The admissible distortions are D_1 for watermark embedding and D_2 for the attacker.

2.1. Capacity for Gaussian channels

The closed-form solution for data-hiding capacity has been found for i.i.d. Gaussian host signal and squared-error distortion functions[6]:

$$C = \Gamma(\sigma_x^2, D_1, D_2) = \begin{cases} \frac{1}{2}\log_2\left(1 + \frac{D_1}{D}\right), \text{ if } D_1 < D_2 < \sigma_x^2 \\ 0, \text{ if } D_2 > \sigma_x^2, \end{cases} \tag{2}$$

where $D = \sigma_x^2 \frac{D_2 - D_1}{\sigma_x^2 - D_1}$. The optimal attack is the cascade of a minimum mean square error (MMSE) estimator for $\mathbf{x}$ given $\mathbf{y}'$, and of a Gaussian test channel from rate-distortion theory. In the case of small distortions ($\sigma_x^2 >> D_1, D_2$), which is common in practical data-hiding applications and assuming large variance, we have $D \sim D_2 - D_1$ and $C = \frac{1}{2}\log_2\left(1 + \frac{D_1}{D_2 - D_1}\right)$, i.e. the capacity expression is *asymptotically independent* of σ_x^2.

2.2. Capacity for parallel Gaussian channels: main results

We assume that the image is decomposed into K channels using some multirate transform. The grouped image coefficients in each channel are assumed to be independent and all samples in the channel have the same variance and are i.i.d. $N(0, \sigma_x^2[k])$. Let $d_1[k]$ and $d_2[k]$ be the distortions introduced in channel k by the data-hider and the attacker, respectively. The channels are assumed to be independent and memoryless. The optimal data-hiding and attacker strategies in each channel are those for a single Gaussian channel with host variance $\sigma_x^2[k]$ and squared-error distortions $d_1[k]$ and $d_2[k]$. The allocation of powers $d_1 = \{d_1[k]\}$ and $d_2 = \{d_2[k]\}$ between channels satisfies the overal distortion constraints:

$$\sum_{k=0}^{K-1} r_k d_1[k] \le D_1, \sum_{k=0}^{K-1} r_k d_2[k] \le D_2, \tag{3}$$

and the inequaliaty constraints

$$0 \le d_1[k], \tag{4}$$

$$d_1[k] \le d_2[k], \tag{5}$$

$$d_2[k] \le \sigma_x^2[k], \tag{6}$$

for $0 \le k \le K-1$. Let $\Gamma(\sigma_x^2[k], d_1[k], d_2[k])$ be the capacity of channel k and $\sum_{k=0}^{K-1} r_k = 1$.

The watermarking capacity for both blind and private parallel-Gaussian watermarking games subject to distortion constraints (D_1, D_2) is equal to[1]:

$$C = \max_{d_1} \min_{d_2} \sum_{k=0}^{K-1} r_k \Gamma(\sigma_x^2[k], d_1[k], d_2[k]) \tag{7}$$

where

$$\Gamma(\sigma_x^2[k], d_1[k], d_2[k]) = \frac{1}{2} r^* \log_2\left(1 + \frac{d_1[k]}{d_2[k] - d_1[k]}\left(1 - \frac{d_2[k]}{\sigma_x^2[k]}\right)\right) \tag{8}$$

where $r^* = \sum_{i=0}^{K^*-1} r_i \in [0, 1]$ is the fraction of strong signal components and the maximization and minimization are subject to the above overall distortion constraints (3-6). The parallel Gaussian watermark channel is schematically shown in Figure 3 for kth channel.

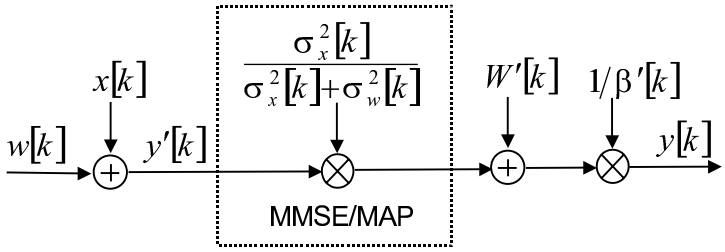


Figure 3. Moulin's optimal data-hiding and attack strategies for parallel Gaussin channels. Each coefficient $x[k] \sim N(0, \sigma_x^2[k])$, $0 \le k \le K-1$. The data hider allocates to each channel watermark with the power $D_1[k]$= $\{\sigma_w^2[k]\}$ resulting it total power D_1. It is assumed that $w[k] \sim N(0, \sigma_w^2[k])$.The attacker intriduces to each channel the distortion $d_2[k]$ resulting in total attacked image distortion D_2. This distortion is distributed between the MMSE/MAP estimator distortion and the test channel distortion. The estimator distortion is equal to the variance of the estimator $\sigma_e^2[k] = \frac{\sigma_x^2[k]\sigma_w^2[k]}{\sigma_x^2[k]+\sigma_w^2[k]}$. All channels are treated independently.

2.3. Capacity for spike process model

Moulin and Mihcak[2] have used so-called *spike process* model introduced by Weidmann and Vetterli.[4] Under a spike model there are two types of channels: those with large variance, i.e. strong channels $\sigma_x^2 >> D_1, D_2$, and those with low variance, i.e. weak channels $\sigma_x^2 << D_1, D_2$. The image components are assumed to be very sparce and independent. Assume that $\sigma_x^2 >> D_1, D_2$, for $0 \le k \le K^* - 1$ and $\sigma_x^2 << D_1, D_2$ for $K^* - 1 \le k \le K - 1$.

The capacity for spike image model is then determined as[2]:

$$C = \frac{1}{2} r^* \log_2 \left(1 + \frac{D_1}{D_2 - D_1}\right). \tag{9}$$

One can make some important conclusions for the capacity based on the spike model.

(1) The capacity does not depend on $\sigma_x^2[k]$ for strong channels.

(2) Almost all energy of the watermark D_1 is allocated to the strong channels and the per-sample capacity is the same for all strong channels.

(3) The attacker cannot considerably attack the strong channels. The extreme case of the attack is to put the strong channels to zero (or to decrease their variance). However, this leads to strong degradations (like in the rate-distortion theory). In the extreme case of zero-rate allocation, it will be equal to the variance of the strong channel $\sigma_x^2[k]$.

The MMSE filter cannot considerebly help kill the watermark in the strong channels, because it does not produce an accurate estimate of the host signal. The variance of the MMSE/MAP estimators in the Gaussian case for both the host image and the watermark is determined as $\sigma_{MMSE}^2[k] = \frac{\sigma_x^2[k]d_1[k]}{\sigma_x^2[k]+d_1[k]}$. According to the spike model assumption for the strong channels $\sigma_x^2[k] \gg d_1[k]$. Therefore, $\sigma_{MMSE}^2[k] \to 1$ and the MMSE estimator term $\frac{\sigma_x^2[k]}{\sigma_x^2[k]+d_1[k]} \to 1$ resulting in $\hat{x}[k] \to y'[k]$ as one-to-one mapping, i.e. no reliable estimate is produced in this case. The only attack that is introduced in this case is a *test channel* from the rate-distortion theory.

3. EXTENDED DATA-HIDING GAMES

The original information-theoretic data-hiding game proposed by Moulin and Mihcak assumes that both the data-hider and the attacker share the same stochastic image model for their max-min strategies. Moreover, a particular form of stochastic image model, i.e. EQ or spike model, is assumed to tackle the sparcity of transform image coefficients. Therefore, the obtained results refer to one of these models. We present here more general set-up that can be valid for many practical applications.

First of all, a natural question concerns the capacity estimate obtained based on the EQ/spike model. Is this estimate the "absolute" limit of data-hiding capacity for a particular image? Secondly, we question existance of

an objective "absolute" capacity limit, since the model selection procedure does not posses an objective character. Thirdly, we tackle the problem of the obtained capacity limit robustness with respect to the variations of stochastic image models. We will follow further the above parallel Gaussian source set-up. The particular interest is to investigate the variability of the obtained capacity estimates for the different image models. Let us first assume we can find the data-hiding capacity C^{IM_1} for image model IM_1 (for example the spike model):

$$C^{IM_1} = \max_{\{d_1, IM_1\}} \min_{\{d_2, IM_1\}} \sum_{k=0}^{K-1} r_k \Gamma(\sigma_x^2[k], d_1[k], d_2[k]), \tag{10}$$

and for some other model IM_2 that is feasible for the given image and provides the memoryless and independent channel decomposition:

$$C^{IM_2} = \max_{\{d_1, IM_2\}} \min_{\{d_2, IM_2\}} \sum_{k=0}^{K-1} r_k \Gamma(\sigma'^2_x[k], d_1[k], d_2[k]), \tag{11}$$

where $\sigma'^2_x[k]$ is the local variance for the model IM_2 assumed to be also i.i.d. locally Gaussian.

The second possible scenario we encompass is based on taking into account *three* strategies, by distinguishing those of the data-hider at the encoder, the attacker and the data-hider at the decoder. We assume that a proper channel state estimation (CSE) is performed that is able to completely characterize the attack channel in terms of geometrical synchronization, fading and noise distribution.[7] Therefore, the data-hider at the decoder is able to use this information to maximize the channel capacity. In fact, this scenario is commonly used in practice in digital communications and has the same practical importance for data-hiding technologies.

Imagine a practical situation where some watermarking technology is developed and maintained during a certain time on the market. The design of this technology assumed that the data-hider and the attacker share the same image model as the solution to:

$$C_{CSE}^{IM_1} = \max_{\{DH_{encoder}, d_1, IM_1\}} \min_{\{Att, d_2, IM_1\}} \max_{\{DH_{decoder}, IM_1\}} \sum_{k=0}^{K-1} r_k \Gamma(\sigma_x^2[k], d_1[k], d_2[k]), \tag{12}$$

where $DH_{encoder}, Att$ and $DH_{decoder}$ denotes the admissible strategies of the data-hider at the encoder, the attacker and the data-hider at the decoder, respectively. A certain amount of images is watermarked using this technology and from the marketing perspective the watermark decoder, designed assuming the image model IM_1, should be maintained on the market during some time.

However, a more realistic scenario is to assume that the attacker is following the recent trends in image processing and possess some more powerful image model IM_2 that outperforms the model IM_1 at least for some reference applications such as image denoising or compression. We formulate here two possible sub-problems, considering either a non-informed or an informed decoder with respect to the image model IM_2 used by the attacker.

The first sub-problem assumes a non-informed decoder, i.e. no side information about the model used by the attacker is available for the decoder:

$$C_{CSE}^{non-informed} = \max_{\{DH_{encoder}, d_1, IM_1\}} \min_{\{Att, d_2, IM_2\}} \max_{\{DH_{decoder}, IM_1\}} \sum_{k=0}^{K-1} r_k \Gamma(\sigma_x^2[k], \sigma'^2_x[k], d_1[k], d_2[k]). \tag{13}$$

Therefore, the joint pair encoder-decoder does not even know which model was used by the attacker, but the decoder still assumes the first model. It is an interesting problem to estimate the loss in performance with respect to this mismatch determined by relative entropy $D(p_{IM_1}||p_{IM_2})) = E_{p_{IM_1}} \log \frac{p_{IM_1}}{p_{IM_2}}$ between two p.d.f.s p_{IM_1} and p_{IM_2} corresponding to IM_1 and IM_2, respectively.

Finally, regarding the second sub-problem, we can assume that the technology provider is also following the recent trends in stochastic image modeling and can integrate a new decoder that assumes both possible models for the attacker. However, an essential amount of images is already watermarked for the model IM_1 and the technology provider should garantee the maintenance of the copyright or watermark detection in general. In this case, the valid formulation is:

$$C_{CSE}^{informed} = \max_{\{DH_{encoder}, d_1, IM_1\}} \min_{\{Att, d_2, IM_2\}} \max_{\{DH_{decoder}, IM_1, IM_2\}} \sum_{k=0}^{K-1} r_k \Gamma(\sigma_x^2[k], \sigma'^2_x[k], d_1[k], d_2[k]). \tag{14}$$

In this paper, we only consider the problem of model impact on the capacity limit according to the set-ups (10) and (11).

4. IMAGE MODEL SELECTION

The model selection is a very important but at the same time a very ambiguous problem. It involves a lot of subjective experience dealing with the estimation, detection and rate-distortion problems. Therefore, this procedure can not be objective or automatic.

This conditions the necessity to justify the fundamental capacity limits for the EQ and spike models. The selection of the EQ or spike models is explained by their excellent performance in some reference applications such as image compression and denoising and good fit to the parallel Gaussian channel model. In a more general case, the advantages of one model over another model are considered based on the satisfaction of a list of requirements which determine the suitability of the model for the practical applications:

(a) model simplicity (preferably Gaussian-type models due to easy integration and differentiation);

(b) model ability to lead to a closed-form analytical solution;

(c) model and result tractability and existance of performance bounds (preferably Gaussian-type models due to the upper and lower bounds in channel capacity and rate-distortion theory);

(d) model robustness and applicability to a wide class of real images.

5. EDGE PROCESS (EP) MODEL

5.1. Model definition

We consider an image $\mathbf{x}$ with a support S as a realization of a random field X with distinct stochastic behavior in different regions. Let $R_1, R_2, ..., R_M$ be a partition of the support S of $\mathbf{x}$, i.e., $R_i \cap R_j = \emptyset, i \neq j$ and $\cup_i R_i = S$. Let $\mathbf{x}_l$ denotes the subset of image pixels supported by the regions R_l. In our model, we assume that each region R_l is fully covered by the model $\theta_l \in \{\Phi_1, \Phi_2, ..., \Phi_L\}$ and that no two neighboring R_l contain the same model. In particular, we assume that the pixels in the image subregion $\mathbf{x}_l$ are distributed with joint probability density function (pdf) $p_{\mathbf{x}}(\mathbf{x}_l \mid \theta_l)$.

Our goal is to introduce the edge process model and to compare it with the EQ model. The EQ model belongs to the class of intraband stochastic image models and assumes that the wavelet coefficients are Gaussian (in the original paper of Lopresto a Generalized Gaussian[3]) distributed, with zero mean and variances that depend on the coefficient location within each subband. It is also assumed that the variance is slowly varying.

We assume that each subband of multiresolution critically sampled transform has its own support $S_l, l = 1, ..., 3M$, where M is the number of diadic decomposition levels such that $S_i \cap S_j = \emptyset, i \neq j$ and $\cup_l S_l = S$. For non-decimated wavelet transform without downsampling used in our modeling, each subband has the same support as above but the dimensionality of each S_l is the same as the original image. According to the partition approach applied to the EQ model, we assume that only one region R_l is given within the subband S_l and all coefficients in this subband belong to the same region R_l:

$$R_l = \{\mathbf{x}_l : x_l[i] \sim N(0, {\sigma_{xl}}^2[i])\} \tag{15}$$

and all coefficients are considered to be independent identically distributed (i.i.d.) with Gaussian distribution, but with different local variances ${\sigma_{xl}}^2[i]$. Equivalently, this means that only one stochastic model out of $\{\Phi_j\}$ is applied to the whole support S_l and $p_{\mathbf{x}l}(\mathbf{x}_l \mid \theta_l)$ follows an i.i.d. Gaussian p.d.f.. In the following, we will consider only the image model for one subband.

Contrarily to the EQ model, the edge process model assumes two distinctive sets of coefficients in wavelet domain for each subband, i.e. those belonging to the flat regions and those belonging to the edge and texture regions. Moreover, it is assumed that a transition corresponding to an edge or to a fragment of texture consists of several distinct mean values that propagate along the transition. In the following we will refer to the transition simply as the edge. These mean values could be considered as the reconstruction levels of the optimal Lloyd-Max scalar quantizer. The variation of the coefficients with the same mean is supposed to be low along the edge. According to the above partition approach, the edge process model is defined as:

$$R_1 = \{\mathbf{x} : x[i] \sim N(0, \sigma_x^2[i])\}, R_2 = \{\mathbf{x} : x_j[i] \sim N(\bar{x}_j[i], {\sigma_x}_j^2[i])\}, \tag{16}$$

where $R_1 \cup R_2 = S$ and S represents a particular subband. The region R_1 represents all flat regions within a subband assumed to be zero-mean Gaussian random variables with the local variance $\sigma_x^2[i]$. The region R_2 corresponds to the texture and edge regions. Each distinctive geometrical structure corresponding to the edge or texture transition within R_2 is decomposed in a set of local mean constellations. Moreover, a particular mean value $\bar{x}_j[i]$, $j = 1, ..., J$ propagates along the edge creating the so-called *edge process* (EP). Therefore, coefficients on the edge are considered to have one of the possible mean values from the set $\{\bar{x}_j[i]\}$, contrarily to the EQ model which does not differentiate flat and edge regions and assumes zero-mean for all coefficients. Moreover, we can also assume that the variation of the coefficients with respect to the mean values (and this is especially true for the overcomplete transform) is very small. Therefore, the EP model assumes that the image consists of random Gaussian process with zero-mean and some small local variance for the flat regions with almost "deterministic" edge occlusions which have a clearly defined geometrical structure depending on the mutual orientation of the edge and the subband. Moreover, normally transitions along the edge have longer stationary length than the transitions within the texture (that explains the existence of higher correlations along the edges); this provides higher redundancy of the support for more accurate model parameter estimation. Due to this fact, the stationarity condition is more strict for the edges than for the textures. Finally, all this leads to the conclusion that the real variance of the subbands is very low and is mostly determined by the flat regions, contrarily to the EQ model or even more for the spike model where the huge spikes of image coefficients with large variance can occur due to the edge that is supposed to model the wavelet coefficients' sparcity. No relationship or special geometrical spatial structure is assumed among the spikes contrary to the EP model where the "spikes" belonging to the same edge are treated jointly along the direction of edge propagation.

5.2. Model parameters' estimation

The goal of this section is to analyze the problem of model parameter estimation and to point out the main drawbacks of the EQ/spike models. We assume that a maximum likelihood (ML) estimation is used to estimate the local variance for the EQ model in some local neighborhood $\Omega(k)$. We assume a LxL square window $\Omega(k)$ centered at location k. The estimation of the local variance according to the ML estimate is:

$$\hat{\sigma}_x^2[i] = \frac{1}{\mid \Omega \mid} \sum_{\kappa \in \Omega(k)} \mid x[\kappa] \mid^2 \tag{17}$$

where $\mid \Omega \mid$ is the cardinality of Ω. Although this ML variance estimator is widely used in the image processing community, its usage is justified only for stationary data which is not the case for the wavelet coefficients. Moreover, the ML variance estimator is only asymptotically unbiased, i.e. $E[\hat{\sigma}_x^2] = \sigma_x^2 - \frac{1}{|\Omega|}\sigma_x^2$. The bias in the variance estimation is $\frac{1}{|\Omega|}\sigma_x^2$ and to reduce the bias one should increase the sampling space Ω, i.e. $\lim_{|\Omega| \to \infty} E(\hat{\sigma}_x^2) = \sigma_x^2$. However, this contradicts to the non-stationary nature of wavelet coefficients. The condition of stationarity is especially violated in the vicinity of edges and textures. As a result, a lot of outliers from the wavelet coefficients with different local means are in the square window and one obtains extremely high estimates of local variance. To avoid this negative effect, one needs to reduce the window size, but this can lead to an increase of the bias. That is why an adaptive window selection based on a bootstrap method was used to achieve a compromise between these two conflicting requirements.[8]

Therefore, the large window was selected for the stationary coefficients in flat regions and the window size was decreased for the edge and texture regions. However, even in this case, the smallest local window of size of 3x3 can not garantee a reliable variance estimate for the edge regions due to critical bias and non-stationary coefficients inclusion especially for the detail subbands which will contain large positive and negative samples even in the smallest possible window size. This situation is schematically explained in Figure 4. Figure 4a represents some edge structure in the transform domain without downsampling and Figure 4b the same edge structure but after downsampling. The ML estimate corresponding to equation (17) is shown in Figure 4c for the edge wavelet coefficient. One can clearly observe the above main problem of the ML estimate due to the non-stationarity of wavelet coefficients. The local window contains both the coefficeints belonging to the edge and the coefficients belonging to the flat region. Contrarily, the EP model (Figure 4d) computes the mean only along the stationary edge that corresponds to the stationarity condition for the ML-estimate, i.e. it takes only those coefficients belonging to the set R_2 for a particular mean constellation.

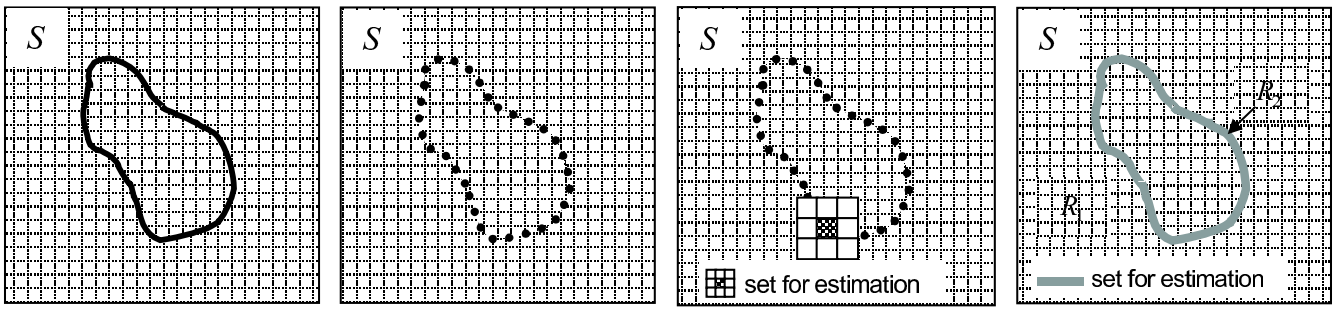


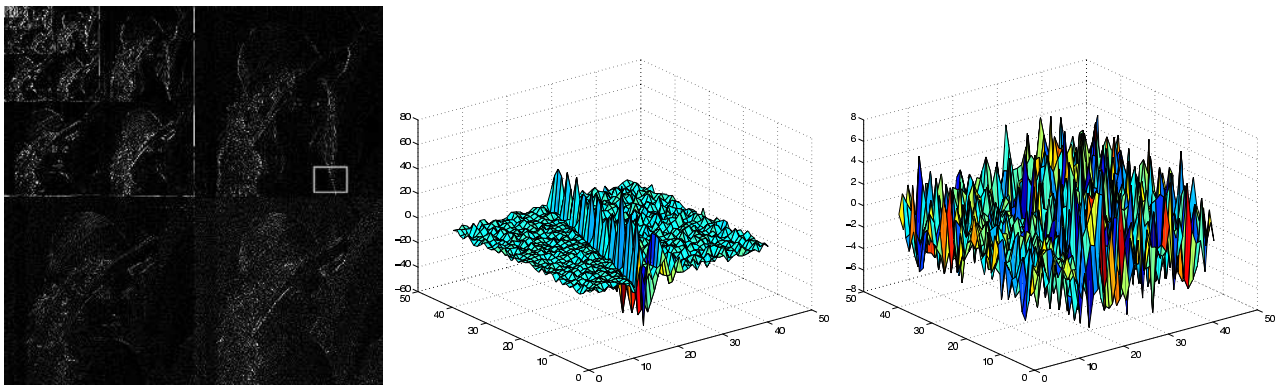
Figure 4. Explanation of the EQ and edge process EP models: (a) the edge structure in the transform domain (overcomplete representation without downsampling); (b) downsampled sparse data in the case of critically sampled transform (wavelets); (c) ML-estimation of the local image variance for the edge coefficient in the local set using a globally stationary assumption; (d) ML-estimation of the local image variance for the edge coefficient using a stationary set of coefficients only along the edge.

5.3. EP model verification on real images

To demonstrate the main features of the EP model on real images we have chosen the image Lena of size 512x512 and decomposed it using a 5-level wavelet transform with *Db*8 filter (Figure 5a). The EP model aims at a more accurate modeling of edges in the transform domain. As an example, we have selected a fragment of the first horizontal subband on the shoulder of Lena image (Figure 5b) that corresponds to a well distinguished propagating edge. One can also observe some variations along the edge due to the critically sampling character of wavelet transform. The edge structure will be smoother, i.e. with smaller variance, for the *overcomplete* transform without downsampling.[9] Small variations of wavelet coefficients are observed in the flat regions on both sides of the edge. The estimation of the mean along the edge according to the EP model, followed by its subtraction, results in the more uniform random field shown in Figure 5c. It is important to note that the amplitude of the coefficients is considerably reduced. The "edge" is not anymore visually detectable and the fragment looks to be more decorrelated.

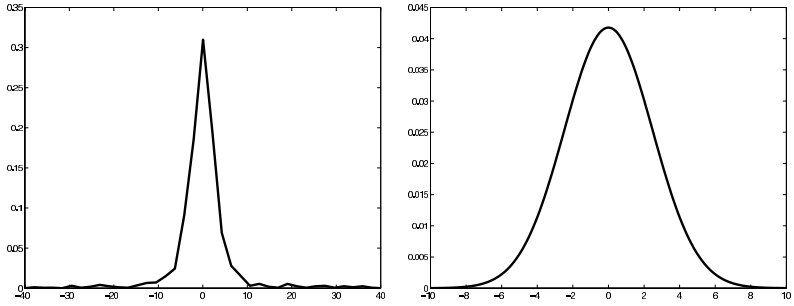
Important changes are also observed for the subband histograms (Figure 6). An essentially non-Gaussian histogram of the original subband (Figure 6a) is transformed into a histogram with perfect Gaussian fit (Figure 6b). Therefore, besides the fact of additional data "decorrelation", the EP model also produces transform coefficients with Gaussian p.d.f.. Since both the EQ and spike models refer to the local variance as the model parameter, we visualize the subband local variances in Figure 7a for the EQ model and in Figure 7b for the EP model. 8-mean constellation was used in the EP model for the R_2 region. Obviously, this simple constellation scheme causes some approximation error that gives rise to an increase in local variance in the vicinity of the edge. More powerful shape approaximation techniques can be used for this purpose and the work is underway. The ML variance estimate was used in both cases. One can also observe a significant decrease of the local image variance, roughly 150 times with respect to the variance peak values. The decrease of the variance has a crucial impact on the performance of denoising and compression algorithms as well as on the capacity estimation problem.

To demonstrate the model ability to carry out and to capture only significant image components with a certain level of sparcity, we performed a set of experiments shown in Figure 8. First, the image was decomposed using *Db*8 wavelets in 5-level pyramid. Second, the EP model was applied and the edges were replaced by their mean estimates according to the EP model. Figure 8a shows the image reconstructed based on the completely preserved information about flat regions and the mean estimates along the edges. Figure 8b is reconstructed only based on the estimate of means and zeros in all remaining positions (total number of non-zero coefficients 66660 out of 262144 for the 512x512 original image). The obtained results demonstrate the high sparsity of the EP model and also the fidelity to the original data. It demontrates the possibility to design new more powerful coders based on shape coding to capture the edge structures.



(a) (b) (c)

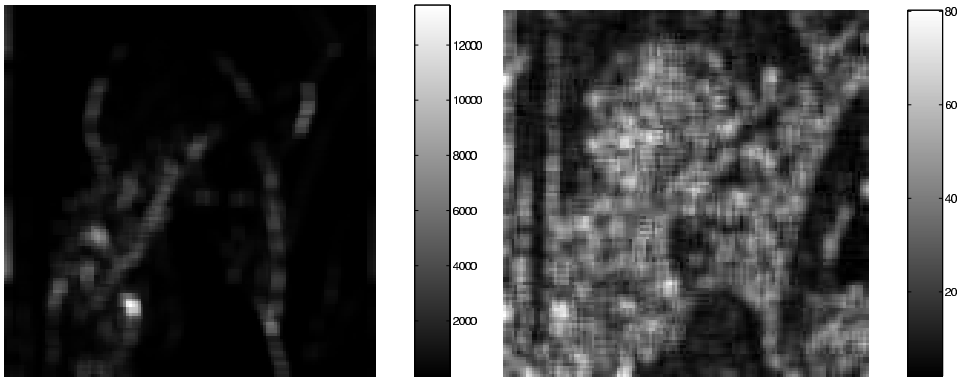
Figure 5. Explanation of the EP model: (a) 5-level orthogonal wavelet transform (*Db*8) applied to image Lena; (b) a subband fragment corresponding to the edge on the Lena's shoulder taken in the first horizontal subband and (c) the same fragment after EP model mean subtraction along the edge.



(a) (b)

Figure 6. Histograms of the first level horizontal subdand before (a) and after EP model mean subtraction (b).

Therefore, we can briefly summarize the main features of the EP image model. First, the EP model offers an additional data "decorrelation" even for the fixed transform basis functions, which could be a very useful feature for many applications such as denoising, compression and watermarking. It should also be noted that the complexity of this "decorrelation" transform, besides model overhead, still remains almost the same as the complexity of the corresponding wavelet or overcomplete transforms. Secondly, the resulting distribution of the subband coefficients is Gaussian as it is demonstrated in Figure 6. This further considerably simplifies the analysis and guaranties the existence of closed-form solutions for many applications, contrarilly to non-Gaussian image models. Since the data is Gaussian and decorrelated this also brings an additional benefit for the independent character of the coefficients. This allows modeling joint subband p.d.f. as a product of independent p.d.f.s of each coefficient. The idea to use the parallel Gaussian channel for the analysis of data-hiding capacity. Thirdly, the subraction of the local mean along the edges makes data "stationary" and considerably reduces the variance estimation based on the ML-estimator. This has an important impact on the performance of denoising and compression algorithms as well as provides a completely different justification for data-hiding algorithm performance opposed to the EQ model (smaller host interference for spread spectrum based watermarking techniques and different capacity estimation limits). In addition, the EP model provides new possibilities for the design of advanced watermarking attacks.



(a) (b)

Figure 7. Variance estimation in wavelet domain for *Db*8 for the second horizontal subband (increased in size for visualization): (a) ML variance estimation in a 5x5 window used in the EQ/spike process model and (b) ML variance estimation for the EP model computed in a 5x5 window. Note that the ranges of variance values decrease from [0..13000] (EQ/spike model) down to [0..80] (EP model).

(a) (b)

Figure 8. Reconstrunction of the Lena image in wavelet domain for *Db*8: (a) from the means of the EP model and completely preserved information about flat regions (PSNR= 38.90dB) and (b) only from the means of the EP model (PSNR= 36.95dB).

Table 1. Comparison of empirical bounds for the EP and EQ models in a denoising application, based on resulting PSNR [dB].

Image model	**Standard deviation of noise**				**Wavelet type**
	10	15	20	25	
	L E N A				
Noisy image	28.13	24.63	22.10	20.17	
EQ (fixed window)	34.44	32.21	30.50	29.18	*Db*8
Mihcak (bootstrap)	34.58	32.59	31.17	30.06	*Db*8
EP (positions)	**34.73**	**32.78**	**31.36**	**30.23**	*Db*8
EP (positions)	**36.21**	**34.44**	**33.00**	**31.83**	9/7, *overcomplete*
EP	**36.52**	**34.81**	**33.55**	**32.53**	9/7, *overcomplete*
	B A R B A R A				
Noisy image	28.14	24.63	22.11	20.18	
EQ (fixed window)	32.97	30.45	28.73	27.38	*Db*8
Mihcak (bootstrap)	32.84	30.46	28.86	27.65	*Db*8
EP (positions)	**33.09**	**30.87**	**29.31**	**28.16**	*Db*8
EP (positions)	**34.83**	**32.83**	**31.41**	**30.23**	9/7, *overcomplete*
EP	**35.05**	**33.16**	**31.92**	**30.78**	9/7, *overcomplete*

5.4. EP model performance in a reference application: denoising

An important aspect of the model selection is its performance in some reference applications. We have chosen the problem of removal of the AWGN from images as a reference application due to its simplicity and existence of benchmarking results. Since, the EQ model is used in the denoising method of Mihcak *et al*[8] we have chosen this method for the sake of comparision. Moreover, to avoid any subjective implementation issues (extraction of edges) of a particular model parameter estimation we have used an *empirical upper bound* as a criterion for the comparison. The empirical upper bound assumes that the stochastic model parameter estimation is performed based on the clean original image. In the case of the EQ model it refers to the estimation of the local image variances for each subband. Moreover, we have investigated two modifications of the EQ models, either using a fixed window for all coefficients, or using a bootstrap version of the EQ denoiser described by Mihcak *et al*.[10] In the case of the EP model, we assume that the information about the edge positions is available and the local variance is directly estimated from the noisy image in the first set of experiments. The second experiment assumes that the local variance is estimated from the original image for fair benchmarking with EQ model. Moreover, to demonstrate the upper limit of the EP model performance, we also performed the denoising in the overcomplete domain. The empirical upper bounds are shown in Table 2 for two test images Lena and Barbara. More results concerning the EP model performance can be found in our paper.[9] Therefore, it is obvious that the EP model outperforms the EQ model in denoising applications by about 0.3-1dB for the critically sampled wavelets and the gap is increased up to 2.8-3.4 dB for the overcomplete domain. This means that not only model selection is important, but also that the transform domain could have a significant impact on the capacity estimation problem. In conclusion, the usage of the EP model as opposed to the EQ model is justified according to this reference application.

6. CAPACITY FOR THE EP MODEL

Assuming the parallel Gaussian channel energy allocation for the spike model, we estimate the data-hiding capacity in the case of the EP model. We use the technique proposed by Moulin and Mihcak[2] for this purpose. The capacity is estimated for the attacker distortion $D_2 = 2D_1$ and for $D_2 = 5D_1$ for several test grayscale images Lena, Barbara and Baboon of size 512x512 (Table 2). The results for the spike model are not surprising. The images with large amount of transitions like Baboon and Barbara are characterized by large $\{\sigma_x^2[k]\}$ and a large embedding distortion D_1 for Baboon image. Thus the capacity estimate is higher than for the rather smooth image Lena. The EP model produces much lower estimates of data-hiding capacity due to a more accurate assumptions about image statistics in

Table 2. Comparison of total data-hiding capacities (in bits) for the test images of size of 512x512, for just noticible distortion D_1 for the spike and EP models.

Image	D_1	$D2 = 2D_1$		$D2 = 5D_1$	
		NC(spike)	NC(EP)	C(spike)	C(EP)
Lena	10	31855	7857	6951	1026
Barbara	10	54632	8480	13045	1237
Baboon	25	80952	2568	17562	436

the transition regions. Since the variance of the EP model is significantly lower in these regions, which are supposed to be the main carriers of the watermark according to the max-min approach and EQ/spike model, the difference in the data-hiding capacity between the spike and the EP model is considerable. One can argue that the capacity for the EQ/spike model is higher and thus it should be considered as an upper bound. The difference between the EQ model capacities in wavelet and DCT domains was explained by the fact that wavelets produce better independence between the cahnnels and better fit to Gaussian distribution.[2] However, comparing the EP and EQ models, as it was shown is Section 5, one can clearly observe that these conditions are even better met for the EP model rather that for the EQ model. Therefore, the existence of larger capacities for the EQ model does not immediately mean that these results should be considered to be tied to practically reachable upper bounds.

7. ATTACK BASED ON EP MODEL

The statistics of real images under the EP model can also be used to develop a more involved attacking strategy. The main idea of this attack is based on the nice approximating feature of the EP model, as demonstrated in Figure 8. We have shown that by only preserving information about means along the edges, one can obtain an image of high quality. Therefore, instead of "killing" the rate of spike coefficients with large variance according to the Gaussian test channel and assumption about zero-mean, we propose to set to zero all flat regions and to replace all edges by the corresponding mean estimates, thus completely killing the watermark in these regions. This attack is an asymptotic case of the Gaussian test channel for zero-rate. In this case, the attack distortion will be proportional to the variance of the edge process, that is relatively small in the case of the EP model contrary to the zero-mean EQ model.

If we assume some edge wavelet coefficient to be a constant value θ along the EP defined edge structure, and that a max-min energy allocation for the i.i.d. watermark $\mathbf{w}$ with energy σ_w^2 is performed, we obtain the model:

$$y[i] = x[i] + w[i] = \theta + w[i]. \tag{18}$$

Assuming the watermark to be i.i.d. Gaussian $w[i] \sim N(0, \sigma_w^2)$, we can apply the ML estimate for θ:

$$\hat{\theta}_{ML} = \frac{1}{\mid \Omega' \mid} \sum_{j \in \Omega'} y[j], \tag{19}$$

where Ω' is a support of the edge and $\mid \Omega' \mid$ is the length of the edge. The ML mean estimate is the unbiased estimate $E[\hat{\theta}_{ML}] = \theta$ with the variance of the estimate $\mathbf{Var}(\hat{\theta}_{ML}) = \mathbf{Var}(\mid \hat{\theta}_{ML} - \theta \mid^2) = \frac{1}{|\Omega'|}\sigma_w^2$. Therefore, the longer the edge $\mid \Omega' \mid$, the more accurate mean estimate one can receive. In the case of an i.i.d. Gaussian assumption about region R_2 (see equation (16)) $x[i] \sim N(\theta, \sigma_x^2)$, the variance of the mean estimator of θ will be $\mathbf{Var}[\hat{\theta}_{ML}] = \frac{\sigma_w^2 + \sigma_x^2}{|\Omega'|}$. In the case of a simple replacement of i.i.d. $x[i] \sim N(\theta, \sigma_x^2)$ by the sample mean estimated from the watermarked image, the resulting variance of the replacement error will consist of $E[\mid \hat{\theta}_{ML} - \mathbf{x} \mid^2] = \sigma_x^2 + \frac{\sigma_w^2 + \sigma_x^2}{|\Omega'|}$ that represents the above mentioned asymptotic case of zero-rate. Although this attack completely kills the watermark in the flat regions and replaces the edge structures by the EP model mean, the inroduced distortion will be smaller than those for the EQ model and Gaussian test channel. This is due to the fact that since one exploits the redundancy along the stationary edge to get more accurate estimate of the mean and lower local variance along the edge (contrarily to the EQ/spike models where all coefficients are treated separately).

8. CONCLUSIONS

We have considered the problem of data-hiding capacity for real images by applying results obtained by Moulin and Mihcak for parallel Gaussian channels. A new stochastic image model has been introduced for the wavelet domain, the so-called EP-edge process model, that more accurately treats data in the regions of edges and textures. We emphasize the crucial role of model selection on determining capacity. Although, the introduced edge process model provides an even more accurate image modeling, that has been proven in a reference application, and fits better the conditions of proper image decomposition in the parallel Gaussian model, the obtained results considerably deviate from those obtained for the EQ/spike models argued to be "upper bounds" on actual capacity. Therefore, the objective character of these bounds for determining the capacity of real images is under question. We have in fact shown that capacity likely to be much lower than previously thought. We also demonstrate the important role of model mismatch in the extended data-hiding games. Finally, a new attack based on the proposed EQ model is presented in the paper.

ACKNOWLEDGMENTS

This paper was partially supported by the Swiss SNF grant No 21-064837.01 "Generalized stochastic image modeling for image denoising, restoration and compression", and by the CTI-CRYMEDA-SA and Interactive Multimodal Information Management (IM2) projects. The authors are thankful to Kivanc Mihcak (Microsoft Research, USA) for helpful and interesting discussions.

REFERENCES

1. P. Moulin, "The role of information theory in watermarking and its application to image watermarking," *Signal Processing, Special Issue on Information Theoretic Issues in Digital Watermarking* **81**(6), pp. 1121–1139, 2001.
2. P. Moulin and M. K. Mihcak, "A framework for evaluating the data-hiding capacity of image sources," *IEEE Trans. on Image Processing* **11**, pp. 1029–1042, September 2002.
3. 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*, pp. 221–230, (Snowbird, Utah, USA), 1997.
4. C. Weidmann and M.Vetterli, "Rate-distortion analysis of spike processes," in *Data Compression Conference*, (Snowbird, USA), March 1999.
5. S. Voloshynovskiy, F. Deguillaume, O. Koval, and T. Pun, "Robust digital watermarking with channel state estimation," *Signal Processing* , 2003. submitted.
6. P. Moulin and J. O'Sullivan, "Information-theoretic analysis of information hiding," *Preprint submitted to IEEE Trans. on Information Theory* , October 1999. available from http://www.ifp.uiuc.edu/ moulin/paper.html.
7. S. Voloshynovskiy, F. Deguillaume, S. Pereira, and T. Pun, "Optimal diversity watermarking with channel state estimation," in *IS&T/SPIE's Annual Symposium, Electronic Imaging 2001: Security and Watermarking of Multimedia Content III*, vol. 4134, pp. 23–27, (San Jose, California USA), 21-26 January 2001.
8. M. K. Mihak, I. Kozintsev, and K. Ramchandran, "Spatially adaptive statistical modeling of wavelet image coefficients and its application to denoising," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, (Phoenix, USA), March 1999.
9. S. Voloshynovskiy, O. Koval, and T. Pun, "Wavelet-based image denoising using non-stationary stochastic geometrical image priors," in *IS&T/SPIE's Annual Symposium, Electronic Imaging 2003: Image and Video Communications and Processing V*, (San Clara, California USA), 20-24 January 2003.
10. M. K. Mihcak, I. Kozintsev, K. Ramchandran, and P. Moulin, "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Signal Processing Letters* **6**, pp. 300–303, December 1999.