Information–Theoretic Analysis of Interscale and Intrascale Dependencies Between Image Wavelet Coefficients *

Juan Liu and Pierre Moulin
University of Illinois at Urbana-Champaign
Beckman Institute, Coordinate Science Lab, and ECE Department
405 N. Mathews Ave., Urbana, IL 61801

Contact Author: J. Liu, tel: (217) 244-1089, fax: (217) 244-8371,
Email: *j-liuf@ifp.uiuc.edu*

Sept. 5, 2000. Revised August 5, 2001

Abstract

This paper presents an information–theoretic analysis of statistical dependencies between image wavelet coefficients. The dependencies are measured using mutual information, which has a fundamental relationship to data compression, estimation, and classification performance. Mutual informations are computed analytically for several statistical image models, and depend strongly on the choice of wavelet filters. In the absence of an explicit statistical model, a method is studied for reliably estimating mutual informations from image data. The validity of the model–based and data–driven approaches is assessed on representative real–world photographic images. Our results are consistent with recent empirical observations that coding schemes exploiting inter– and intra– scale dependencies alone perform very well, whereas taking both into account does not significantly improve coding performance. A similar observation applies to other image processing applications.

Keywords — mutual information, rate–distortion, image restoration, image compression, image modeling, wavelets, Markov processes.

EDICS Categories : 2–WAVP, 2–MODL

*Work supported by NSF CAREER award MIP-97-32995 and NSF grants MIP-97-07663 and CDA-9624396. Part of this work was presented at ICIP'00.

1 Introduction

In image processing applications such as compression, estimation, and classification, one can construct optimal or near-optimal algorithms based on accurate statistical image models [1, 2]. For instance, Shapiro's zerotree coding technique [3] has led to a new generation of powerful wavelet image coders that exploit the clustering of wavelet coefficients in scale and space [4, 5, 6, 7, 8]. In image denoising problems, adaptive wavelet filtering techniques [9, 10, 11, 12] and estimators based on hidden Markov models (HMT) [13, 14] outperform simpler Wiener filtering techniques. The use of hidden Markov tree models has also been beneficial in image classification [15, 16]. Markov random field models [17, 18] have been used successfully in some applications.

The potential advantages of using a particular model can be validated by an improved performance in a specific application, as in the above papers, or by a direct characterization of the discrepancy between this model and a simpler one. This paper presents such a characterization of interscale and intrascale dependencies between image wavelet coefficients.

Such dependencies have been studied intensively in the image compression and estimation literature. They can be formulated explicitly (e.g., [5], [6], [8]–[14]), or implicitly (e.g., [4, 7]). The resulting wavelet models can be loosely classified into three categories: those exploiting interscale dependencies, those exploiting intrascale dependencies, and those exploiting both. It is not always clear which type of model should be preferred, and why it should be preferred.

Current image compression and estimation practice suggests that models combining both inter– and intra–scale dependencies models *are not significantly better* than models exploiting intrascale dependencies alone. For example, the recently developed JPEG–2000 image compression standard exploits intrascale dependencies alone [19, 20]. In addition, recent image denoising experiments [21] have compared the performance of a composite model–based estimation scheme and an intrascale estimation scheme. The mean–squared error (MSE) using the first scheme is only slightly ($< 5\%$) lower. This paper seeks an analytical explanation for such empirical observations and develops a framework for studying related questions.

The main theme of this paper is to compare various wavelet models based on their ability to capture dependencies between wavelet coefficients, rather than their experimental performance in any *specific algorithm* measured using an application–dependent criterion such as MSE or compres-

sion ratio. The dependencies between coefficients are measured using mutual information, which has a fundamental relationship to compression, estimation, and classification performance, e.g., in the form of performance bounds.

1.1 Overview of statistical wavelet models

The wavelet transform nearly decorrelates many images and can be viewed as an approximate Karhunen–Loève transform (KLT) [22]. This is the basic property exploited by early wavelet coders and wavelet denoising algorithms. Nevertheless, significant dependencies still exist between wavelet coefficients. Each statistical wavelet model in the literature focuses on a certain type of dependencies, which it attempts to capture using a relatively simple and tractable model. We classify these models in the following three categories.

1. Interscale models. The magnitudes of wavelet coefficients in typical images are strongly correlated across scales. Consider a quadtree representation of wavelet coefficients. If a parent node has small magnitude, its children are very likely to be small too. This property is exploited in Shapiro's embedded zerotree wavelet (EZW) coder [3]. Combining the self–similarity across scales with a clever scheme for set partitioning in hierarchical trees (SPIHT), Said and Pearlman developed an even better coder [4]. The hidden Markov tree model (HMT) by Crouse *et al* [13] also captures the dependencies between a parent and its children. A hidden state is associated with each wavelet coefficient; conditioned on their hidden states, the coefficients are Gaussian, independent and identically distributed (iid).

2. Intrascale models. Strong dependencies in the form of spatial clusters exist between wavelet coefficients inside each subband. Compression algorithms such as the morphological coder [7] exploit the spatial clustering of these wavelet coefficients. The EQ coder [6] models wavelet coefficients as independent generalized Gaussian distributed (GGD) with zero mean and *slowly varying* variance. Local statistics are estimated from the data. This model has recently found applications to denoising [12].

3. Composite dependency models. Both types of dependencies above may be combined. For instance, Joshi *et al* [5] and Liu and Moulin [21] developed classification–based models involving both interscale and intrascale dependencies. Predictive models have been used by Chang *et al* [11]

and Simoncelli [9]. In particular, Simoncelli [9] assumes a strong correlation between the squared magnitude (energy) of a wavelet coefficient and those of its parent and neighbors, and develops a prediction scheme based on that assumption.

1.2 Organization of this paper

Sec. 2 of this paper formulates the modeling problem in terms of mutual information, relative entropy, and Markovian properties. Sec. 3 illustrates these concepts in the special, classical case of an AR–1 Gaussian image process. We compute the mutual informations analytically, and study the influence of the choice of the wavelet filter. In Sec. 4, we move on to more complex, nonparametric models. Modeling and estimating high–dimensional probability density functions (pdf) for neighborhoods of wavelet coefficients is difficult, so we propose a technique to "summarize" the neighborhood based on sufficient statistics. The choice of the summary function is discussed for nonlinear Markov models and doubly stochastic models. In Sec. 5, we describe numerical methods for nonparametric estimation of mutual informations from image data. Results on real–world photographic images are presented and interpreted. In Sec. 6, we consider doubly stochastic models such as the EQ model and derive an upper bound on intrascale mutual information. This yields additional insights about wavelet models and provides a convenient alternative to the numerical mutual information estimation methods in Sec. 5. Discussions are presented in Sec. 7.

2 Mutual Information

To compare interscale, intrascale, and composite wavelet models based on their ability to capture dependencies between wavelet coefficients, we seek a simple but useful quantitative measure of dependency. While a correlation coefficient is appropriate for Gaussian–distributed data, it is typically inadequate for non–Gaussian distributions. For instance, the correlation of adjacent wavelet coefficients within a subband is typically very low (approximately 0.1 [23]), yet inspection of the magnitudes of these coefficients immediately reveals strong dependencies.

Consider mutual information, which admits direct data compression and classification interpretations [24] as well as an estimation interpretation [25, 26]. Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be two

random variables (or vectors) having a joint pdf $p(x,y)$. The mutual information between X and Y is defined as

$$I(X;Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \, dxdy \stackrel{\triangle}{=} E_{XY}\left[\log \frac{p(x,y)}{p(x)p(y)}\right] = D\left(p(x,y)||p(x)p(y)\right), \quad (1)$$

where $D(\cdot||\cdot)$ is the relative entropy between two distributions, also known as the Kullback–Leibler divergence [24]. Throughout the paper, we use logarithm of base 2, hence $I(X;Y)$ is measured in bits. If the differential entropies $h(X) = E_X\left[-\log p(x)\right]$ and $h(X|Y) = E_{X,Y}\left[-\log p(x|y)\right]$ are finite, then $I(X;Y) = h(X) - h(X|Y)$. The mutual information is symmetric in X and Y, nonnegative, and is equal to zero if and only if X and Y are independent. If X is a function of Y, $I(X;Y) = \infty$ [27].

The mutual information $I(X;Y)$ indicates how much information Y conveys about X. For instance, $I(X;Y)$ admits a well-known data compression interpretation: encoding X to a precision Δ_X costs $h(X) - \log \Delta_X$ bits (assuming sufficiently small Δ_X), but if Y is known, encoding X to the same precision given Y costs only $h(X) - \log \Delta_X - I(X;Y)$ bits [24]. The saving is $I(X;Y)$ bits. The conditional mutual information $I(X;Y|Z) \stackrel{\triangle}{=} E_{XYZ}\left[\log \frac{p(x,y|z)}{p(x|z)p(y|z)}\right] = h(X|Z) - h(X|Y,Z)$ admits a similar interpretation.

In estimation problems, mutual information provides bounds on parameter estimation performance via the distortion-rate bound [25, 26]. The higher $I(X;Y)$ is, the easier it is to estimate X given Y or vice-versa. Note that Fisher information also provides bounds on the variance of unbiased estimators via the Cramer-Rao bound, but extension of these bounds to the case of biased estimators and/or non-Euclidean parameter sets is quite unwieldy [28, 29]. Moreover, the Cramer-Rao bound is local and typically tight only for large-sample problems.

Mutual information can be used for adaptive or off–line processing in various applications. Examples include image registration [30, 31], independent component analysis (ICA) [32], and the application of ICA to various problems such as blind source separation [33] and image restoration [34].

In order to better understand the performance of image processing algorithms that exploit interscale dependencies, intrascale dependencies, or both, we refer to Fig. 1 and compare the following mutual informations:

- $I(X;\mathcal{P}X)$, where X denotes a wavelet coefficient, and $\mathcal{P}X$ denotes its parent in the next coarser subband.

- $I(X;\mathcal{N}X)$, where $\mathcal{N}X$ is a predefined neighborhood of X (excluding X). For backward–adaptive coders such as in [6], one is interested in causal neighborhoods, which are used to adapt the quantizer applied to X. For forward-adaptive coders [35], adaptation is done using larger, noncausal neighborhoods.

- $I(X;\mathcal{P}X,\mathcal{N}X)$, corresponding to the composite dependency model which takes into account both parent and neighborhood data $(\mathcal{P}X,\mathcal{N}X)$.

From the chain rule for mutual information [24], we know that $I(X;\mathcal{P}X,\mathcal{N}X) \geq I(X;\mathcal{N}X)$, where the difference between the two terms is $I(X;\mathcal{P}X|\mathcal{N}X)$. This difference quantifies *how much information $\mathcal{P}X$ conveys about X if $\mathcal{N}X$ is already known.* It is zero if and only if $\mathcal{P}X \rightarrow \mathcal{N}X \rightarrow X$ forms a Markov chain. Similarly, we have $I(X;\mathcal{P}X,\mathcal{N}X) \geq I(X;\mathcal{P}X)$. The difference between these two terms, $I(X;\mathcal{N}X|\mathcal{P}X)$, quantifies how much information $\mathcal{N}X$ conveys about X if $\mathcal{P}X$ is already known. It is equal to zero if and only if $\mathcal{N}X \rightarrow \mathcal{P}X \rightarrow X$ forms a Markov chain.

In many applications, one would like the mutual informations listed above to be small, meaning that the wavelet transform nearly whitens the image data. But in fact there are residual dependencies between wavelet coefficients, which can be quantified using mutual information and exploited using appropriate techniques.

3 Special Case: AR–1 Gaussian Model

Assume here that the image is a stationary AR–1 Gaussian process. This is a simple but oft–used model in image processing [36]. In this case, the mutual informations of Sec. 2 can be computed in closed form, and insightful results are obtained.

We use a 1–D signal to simplify the notation and illustrate the basic concepts. Suppose the signal is a Gaussian random process $\{g(n), n \in \mathbb{Z}\}$ with autocorrelation sequence $R_G(k) = r^{|k|}\sigma^2$, where $0 \leq r \leq 1$ and $k \in \mathbb{Z}$. We consider a two–level wavelet decomposition using lowpass filter h_0 and highpass filter h_1, as plotted in Fig. 2. The wavelet decomposition produces three subbands:

the coarse subband B_0, the first fine subband B_1, and the finest subband B_2. For any coefficient in B_2, we compute the correlation with its parent (in B_1) and neighbors (in B_2). The correlation can be expressed as the convolution of $R_G(n)$ (or its down–sampled version) with the filterbank coefficients. Specifically, let $f_1(n)$ denote the filter with z-transform $F_1(z) = H_0(z)H_1(z^2)$. For any $j, h \in \mathbb{Z}$, we have

$$\begin{aligned} E[B_2(j)B_2(h)] &= \sum_l \sum_n h_1(l)h_1(n)r^{|2(h-j)+l-n|}\sigma^2; \\ E[B_1(j)B_2(2j+h)] &= \sum_l \sum_n f_1(l)h_1(n)r^{|2h+l-n|}\sigma^2; \\ E[|B_1(j)|^2] &= \sum_l \sum_n f_1(l)f_1(n)r^{|l-n|}\sigma^2. \end{aligned}$$

Now recall that the mutual information between two Gaussian random vectors X and Y is given by [24]

$$I(X;Y) = \frac{1}{2}\log\left(|R_{XX}| \cdot |R_{YY}| \cdot \begin{vmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{vmatrix}^{-1}\right), \tag{2}$$

where $|\cdot|$ denotes the determinant of a matrix, and R_{XX}, R_{YY}, and R_{XY} are the autocorrelation of X, autocorrelation of Y, and the cross correlation between X and Y, respectively. Define X as the coefficient $B_2(j)$, its parent $\mathcal{P}X$ as $B_1(\lfloor \frac{j}{2} \rfloor)$, and its neighborhood $\mathcal{N}X$ as $\{B_2(j-1), B_2(j+1)\}$. The mutual informations $I(X;\mathcal{P}X)$ (interscale), $I(X;\mathcal{N}X)$ (intrascale), and $I(X;\mathcal{P}X,\mathcal{N}X)$ (composite) can then be computed from (2).

For brickwall filters, we have

$$I(X;\mathcal{P}X) = 0 \quad \text{and} \quad I(X;\mathcal{N}X,\mathcal{P}X) = I(X;\mathcal{N}X).$$

This is because the subbands B_1 and B_2 correspond to different frequency components that are statistically independent under our stationary model. In this case, intrascale models capture all the dependencies.

Fig. 3 plots the mutual information values using Haar filters and Daubechies' maximally flat 8–tap filters [37], as a function of the correlation coefficient r. The value of mutual information does not depend on σ^2, as can be easily verified from (2).

The value of mutual information strongly depends on the filterbanks $\{H_0(z), H_1(z)\}$. In particular, interscale dependencies are significant for the Haar wavelet and weak for the Daubechies wavelet.

Moreover, the Haar wavelet produces larger values of $I(X;\mathcal{N}X,\mathcal{P}X)$ compared to the Daubechies wavelet (e.g., for $r = 0.9$, $I(X;\mathcal{N}X,\mathcal{P}X) = 0.093$ and 0.030 bits, respectively). This is due to the fact that Daubechies' wavelet does a better job of approximately whitening the spatial domain AR–1 process. For a perfectly whitening transform[1], we would have $I(X;\mathcal{N}X,\mathcal{P}X) = 0$. Most of the dependencies remaining after application of the wavelet decomposition using Daubechies' 8–tap filters can be captured by an appropriate intrascale model.

4 Reduced-Order Models

This section is concerned with more complex, possibly nonparametric models for the joint distribution of $(X;\mathcal{N}X,\mathcal{P}X)$. One major practical difficulty in estimating the mutual informations $I(X;\mathcal{N}X)$ (intrascale) and $I(X;\mathcal{N}X,\mathcal{P}X)$ (composite) is the high dimensionality of the models, due to the possibly large size of the neighborhood $\mathcal{N}X$. For example, consider an image subband and define the neighborhood $\mathcal{N}X$ as the collection of the eight coefficients adjacent to X, as shown in Fig. 4. It is difficult to reliably estimate the 9–dimensional pdf $p(x,\mathcal{N}x)$ because the number of data needed to accurately estimate a pdf increases exponentially with the dimensionality [39]. To avoid this so–called curse of dimensionality, we would like to assume that the neighborhood $\mathcal{N}X = \{\mathcal{N}X_1,...,\mathcal{N}X_k\}$ provides information to X only through a many–to–one scalar function $T = f(\mathcal{N}X)$, in the sense that

$$I(X;\mathcal{N}X) = I(X;T). \tag{3}$$

As we shall see in Secs. 4.1 and 4.2, several models considered in recent image processing literature use this assumption. Under (3), the original problem is then reduced to estimating mutual information for a 2–D density, which is a much simpler problem. The assumption that the function $f(\cdot)$ "summarizes" the neighborhood information is illustrated in Fig. 5.

Of course, Assumption (3) is not necessarily satisfied by the actual image model. Regardless of the choice of the function $f(\cdot)$,

$$X \to \mathcal{N}X \to T = f(\mathcal{N}X) \tag{4}$$

[1]Recall that under standard regularity conditions, the Fourier transform whitens stationary, discrete–time, random processes [36][38, Sec. 4.4].

forms a Markov chain, and the data–processing theorem [24, Ch. 2.8] implies

$$I(X;\mathcal{N}X) \geq I(X;T). \tag{5}$$

Equality is achieved if and only if the statistic $T = f(\mathcal{N}X)$ is sufficient for X [24, Ch. 2.10].

4.1 Nonlinear Markov models

Because Assumption (3) rarely holds exactly in practice, finding the proper form of the summarizing function $f(\cdot)$ is important. The inequality (5) suggests a criterion for choosing $f(\cdot)$: maximize $I(X;T)$. This would yield the best model in the mutual-information sense among the class of 2–D models for $(X,\mathcal{N}X)$.

Note that several empirical choices for $f(\cdot)$ have been proposed in the image processing literature [8, 9, 11]. For example, Simoncelli [9] assumes that the squared magnitude of the current coefficient can be linearly predicted from that of its neighbors at the same scale. His model is a nonlinear Markov model which can be stated as follows. The statistic T is a weighted average of $\{|\mathcal{N}X_i|^2\}$, and X is Gaussian, conditioned on T:

$$T = f(\mathcal{N}X) \quad = \quad \sum_i W_i |\mathcal{N}X_i|^2, \tag{6}$$

$$X/\sqrt{T+\sigma_e^2} \quad \sim \quad N(0,1) \text{ is independent of } T. \tag{7}$$

Here σ_e^2 denotes variance of the prediction error in the subband. The model (6) (7) has been validated in [9] using the histogram of the normalized wavelet coefficients $X/\sqrt{T+\sigma_e^2}$ and their correlation. This suggests that the dependencies are mostly captured by the function $f(\cdot)$ defined in (6), and that Assumption (3) approximately holds.

To illustrate the problem of selecting $f(\cdot)$ according to our mutual–information criterion, consider the function (6), and define the neighborhood $\mathcal{N}X$ to be the set of eight coefficients adjacent to X, as in Fig. 4. Two choices of $\{W_i\}$ are considered:

- Equal weights $W_i = \frac{1}{8}$ (see Fig. 4a). Here T is an unbiased estimate of the variance of X (assuming $\{\mathcal{N}X_i\}$ are zero-mean and identically distributed).

- Adaptive weights with W_i's estimated based on the data. This choice allows some flexibility in the choice of $f(\cdot)$. To reduce the number of free parameters, assume symmetric weights, as illustrated in Fig. 4b. The upper, lower, left and right neighbors of X are assigned equal weights W_1, and the four diagonal neighbors are assigned equal weights W_2. Since the mutual information is invariant to linear scaling, $I(X;T)$ depends on W_1 and W_2 only through the ratio $\beta = \frac{W_2}{W_1}$. In this case, we let

$$T = \sum_{i \in \{hor, ver\}} |\mathcal{N}X_i|^2 + \sum_{i \in \{diag\}} \beta |\mathcal{N}X_i|^2, \tag{8}$$

  where *hor*, *ver*, and *diag* denote the two horizontal, two vertical, and four diagonal neighbors, respectively. The parameter β can be estimated from the data, for instance, by maximizing the log–likelihood $\sum_{i=1}^{N} \log p(x_i|t_i)$, where N is the number of available samples. In the asymptotic case $N \to \infty$, the normalized sum $\frac{1}{N}\sum_{i=1}^{N} \log p(x_i|t_i)$ converges in probability to its expectation $E[\log p(X|T)] = -h(X|T)$ according to the weak law of large numbers [40]. Moreover, $-h(X|T) = I(X;T) - h(X)$. Thus the estimate β asymptotically maximizes the mutual information $I(X;T)$, and the inequality (5) is tighter than the bound corresponding to $\beta = 1$ (equal–weights design).

4.2 Doubly stochastic models

Consider a class of models that is analytically and computationally simpler than (6) (7) and has been used in recent image processing literature [6, 12, 41]:

> X and its neighboring coefficients in $\mathcal{N}X$ are independently drawn from a distribution $p(\cdot|\theta)$ parameterized by θ, and θ itself is a random variable following a distribution $p(\theta)$.

In this case, the optimal summarizing function $T = f(\mathcal{N}X)$ is a sufficient statistic for estimating θ from $\mathcal{N}X$, provided such a 1-D sufficient statistic exists [28]. By the description of this model,

$$X \to \theta \to \mathcal{N}X \to T \tag{9}$$

forms a Markov chain. Doubly stochastic models have been used, for instance, in the EQ coder [6] and in the denoising algorithms of Mıhçak *et al.* [12] and Liu and Moulin [21]. Given the local

variance θ, the coefficients are assumed to be independent with distribution $N(0, \theta)$. In this special case, the optimal summarizing function is the minimal sufficient statistics for θ: $T = \sum_i \mathcal{N}X_i^2$.

Another direct consequence of the Markov property (9) is

$$I(X;\theta) \geq I(X;\mathcal{N}X) \geq I(X;T) \tag{10}$$

with equality if and only if T is a sufficient statistic for θ, and θ can be estimated *exactly* from T. This equality is approximately satisfied if the neighborhood is sufficiently large, and if T is a consistent estimator of θ. Under these assumptions, $I(X;\mathcal{N}X)$ can be approximated using a simple parametric model for the wavelet coefficient dependencies.

5 Nonparametric Estimation of Mutual Information

For most distributions, mutual information cannot be computed analytically. Moreover, the pdf's themselves are rarely available. In this section, we develop numerical methods to estimate mutual information based on available wavelet coefficient data. This problem is at least as difficult as pdf estimation; see [42, 43] for estimation of entropy in a general, theoretical context.

5.1 Nonparametric estimators

Given two random vectors X and Y with known joint pdf $p(x, y)$, $I(X;Y)$ is defined by the integral (1). We let X be the wavelet coefficient, and Y be the neighborhood statistic $T = f(\mathcal{N}X)$, the parent $\mathcal{P}X$, or the vector $(T, \mathcal{P}X)$. The pdf $p(x, y)$ is unknown and must be estimated from empirical data. Consider a nonparametric approach. Partition the range of X and Y into N_X and N_Y intervals, respectively. The histogram of (X, Y) obtained from the binned empirical data is denoted as $\{P_{X,Y}(i, j)$, for $1 \leq i \leq N_X, 1 \leq j \leq N_Y\}$ and yields an approximation to the pdf $p(x, y)$. Likewise, the marginal distributions $P_X(i)$ and $P_Y(j)$ can be estimated. From these histograms, the mutual information is estimated as

$$\hat{I}(X;Y) = \sum_i \sum_j P_{X,Y}(i,j) \log \frac{P_{X,Y}(i,j)}{P_X(i)P_Y(j)}. \tag{11}$$

Assume that the random sequence $\{(X_n, Y_n), 1 \leq n \leq N\}$ is stationary and ergodic. Then the histogram yields a reliable estimate of the pdf, and $\hat{I}(X;Y)$ converges to $I(X;Y)$ in probability.

The distribution may vary from subband to subband, but is assumed stationary and ergodic within each subband.

In our experiments, we have used two different binning techniques to construct the histograms used in (11): a log–scale histogram method and an adaptive partitioning method [44]. Both techniques are described in Appendix A.

5.2 Experimental results

We have used a database containing ten representative images, each of size 512×512, ranging from simple (such as *Peppers*) to complicated (such as *Baboon*). The thumbnail version of these images can be viewed at the website http://www.ifp.uiuc.edu/~j-liuf/thumbnails/images.html. Mutual information is computed for these images using the two numerical methods in Appendix A. Due to space limitations, we report the results for representative images such as *Lena*, *Barbara* and *Peppers*.

We used Daubechies' maximally flat 4–tap filters [37] in a 4–level wavelet decomposition. Table 1 reports results computed using the log–histogram method for the finest subbands of the images *Lena*, *Barbara*, and *Peppers*. Two special cases of the mapping $f(\cdot)$ in (6) are compared. The first one assigns equal weights to all neighbors, and the second adjusts the ratio $\beta = \frac{W_2}{W_1}$ in each subband so as to maximize the estimated intrascale mutual information, as described in Sec. 4.1. The estimated β values are listed in Table 1. We notice that the model $f(\cdot)$ with adjustable β is better in the sense that it better captures intrascale and composite model dependencies (see theoretical justification in Sec. 4). However, the improvement is minor (about 2% for *Lena*, and 3% for *Barbara* and *Peppers*. This indicates that despite its simplicity, an equally weighted combination of neighbors yields a good model for intrascale dependencies among wavelet coefficients. If the stationary ergodic assumption of Sec. 5.1 holds and if the assumption (4) holds, we have $\hat{I}(X;T) \approx I(X;\mathcal{N}X)$, where the discrepancy is small and is only due to finite sample size. Likewise, $\hat{I}(X;T,\mathcal{P}X) \approx I(X;\mathcal{N}X,\mathcal{P}X)$.

Table 2 reports the mutual information computed using the log–histogram estimation method for the next–to–finest subbands. The most striking result in Tables 1 and 2 is that $\hat{I}(X;T,\mathcal{P}X)$ is always *significantly larger* (84% for *Lena*, 201% for *Barbara*, and 133% for *Peppers*) than $\hat{I}(X;\mathcal{P}X)$,

and only *a few percent larger* than $\hat{I}(X;T)$ (13% for *Lena*, 6% for *Barbara*, and 13% for *Peppers*). This is also true for other images in the database; see Table 3 for the summary of results. This indicates that *intrascale models capture most of the dependencies between wavelet coefficients, and the gains obtained by also including the parent information are marginal.* This is consistent with the performance of image processing algorithms reported in compression and estimation literature (see Sec. 1).

We also experimented with Haar filters and with Daubechies' maximally–flat 8–tap filters [37]. We observed that

$$\hat{I}(X;\mathcal{P}X) < \hat{I}(X;T) < \hat{I}(X;T,\mathcal{P}X)$$

regardless of the choice of wavelet. For long filters, $\hat{I}(X;T)$ is close to $\hat{I}(X;T,\mathcal{P}X)$, and $\hat{I}(X;\mathcal{P}X)$ is small. This is true for all finest and next–to–finest subbands of our test images. We take the finest horizontal subband of *Lena* as an example. Using a Haar wavelet, $\hat{I}(X;\mathcal{P}X)$ is about 67% of $\hat{I}(X;T,\mathcal{P}X)$, and $\hat{I}(X;T)$ is about 87% of $\hat{I}(X;T,\mathcal{P}X)$; using a Daubechies 4–tap filterbank, the two percentages are 55% and 91% respectively (see Table 1); and using a Daubechies 8–tap filterbank, the two percentages are 49% and 93% respectively.

Table 4 compares our two methods for estimating mutual informations: the log scale histogram method and the adaptive partitioning method [44]. The two methods produce consistent values of the mutual information within about 10%. This suggests that these estimated values are relatively reliable. It is also interesting to see what the partitioned cells using Darbellay and Vajda's method look like. Fig. 6 shows an example plotted in log scale. The range of $(X;\mathcal{P}X)$ is partitioned into a set of cells. The partition is nonuniform and nonseparable (see Appendix A). Interestingly, the resulting partitioning is fairly close to uniform discretization of $\log|X|$ and $\log|\mathcal{P}X|$.

5.3 Discussion on experimental results

Based on these results, one may come to the conclusion that intrascale models should always be favored over interscale models. However, this need not be so. For instance, the intrascale mutual information depends on the size and shape of the neighborhood. In compression applications, backward–adaptive intrascale models such as the EQ coder [6] use a small causal neighborhood (which is also available at the decoder) to help encoding the current coefficient. Consider a causal

neighborhood $\mathcal{N}X$ defined as the upper, left, and upper–left neighbors of X (three coefficients only), and repeat the experiments of Sec. 5.2. The corresponding mutual informations produced are reported in Table 5 for the finest subbands of *Lena*. Compared with Table 1, the advantage of intrascale models over interscale ones is much less prominent.[2]

Moreover, interscale models also have some advantages. Interscale coders such as EZW [3] and SPIHT [4] can be considered as vector quantizers, where the coefficients in a hierarchical tree are classified jointly. The tree structure is algorithmically more convenient than a noncausal neighborhood structure. This is analogous to the difference between a Markov chain and a Markov random field, with the latter being more complicated than the former, both in theory and in practice. Besides the rate–distortion performance of coders, also of concern are practical issues such as the complexity of encoding and decoding, functional requirements such as embedding [3, 4], and hardware implementability.

Related to our comparison of interscale, intrascale, and composite models is the work by Buccigrossi and Simoncelli [8] which linearly predicts a coefficient's magnitude from a conditioning coefficient set. The set may include one or some of the following: the coefficient's parent, neighbors (left and upper), cousins (coefficients at the same location but in different orientation subbands), and aunts (cousins of the parent). To determine which candidates to include in the prediction set, a greedy algorithm compares the mutual information between the coefficient's magnitude and its linear estimator, and includes the most informative candidate in the conditioning set first. The parent is ranked third, providing less information content than the left and upper neighbors. This result is complementary to ours. It also verifies that the dependencies between neighboring coefficients (intrascale) are stronger than the interscale dependencies.

6 Upper Bound $I(X;\theta)$ Under Doubly Stochastic Models

Doubly stochastic models (or more generally, hierarchical models [45]) have been used in image processing applications (e.g., [46]). Recently they have been used to model wavelet domain images [12, 47, 48]. Consider the doubly stochastic model from Sec. 4.2, where the wavelet coefficients

[2]Furthermore, in compression applications, X is encoded given a quantized version of $\mathcal{N}X$ only. According to the data processing theorem [24], such quantization further reduces mutual information.

are drawn independently from a distribution $p(\cdot|\theta)$ parameterized by θ, and the distribution of θ is $p(\theta)$. Recall from (10) that $I(X;\theta)$ gives an upper bound on the intrascale mutual information $I(X;\mathcal{N}X)$, and that this bound is tight for sufficiently large neighborhoods. In this section, we derive mutual informations under several useful models for $p(\cdot|\theta)$ and $p(\theta)$. Our derivation uses the property that mutual information is invariant to invertible mappings such as linear scaling: $I(\alpha X;\xi Y) = I(X;Y)$ for any $\alpha, \xi \neq 0$. The mutual information $I(X;\theta)$ depends on $p(\theta)$, but in some cases is invariant to parameters of $p(\theta)$. Consider the following examples.

- **Model 1** [12, 49]:

$$\begin{cases} p(x|\theta) & \sim & N(0,\theta); \\ p(\theta) & = & ae^{-a\theta} \quad \text{for} \quad \theta \geq 0. \end{cases} \tag{12}$$

The marginal distribution of X is Laplacian with zero mean and variance $\frac{1}{a}$ [49]. To compute $I(X;\theta)$, consider the scaled random variables $\theta' = a\theta$ and $X' = \sqrt{a}X$. The joint distribution of (X', θ') is given by

$$\begin{cases} p(x'|\theta') & \sim & N(0,1); \\ p(\theta') & = & e^{-\theta'} \quad \text{for} \quad \theta' \geq 0 \end{cases}$$

and is independent of a; hence so is $I(X';\theta')$. Due to the scaling–invariant property of mutual information,

$$I(X;\theta) = I(X';\theta') = \frac{1}{2}(1 + \gamma - \log \pi) \text{ nats } = 0.313 \text{ bits},$$

where $\gamma = \lim_{m\to\infty} \left(\sum_{k=1}^{m} \frac{1}{k} - \log m\right) \approx 0.577216$ is Euler's constant. The fact that $I(X;\theta)$ is independent of a is remarkable: as long as θ follows an exponential distribution, $I(X;\theta) = 0.313$ bits; and the variance of the distribution $p(\theta)$ does not matter.

We further generalize this analysis to the case where the variance parameter θ follows a single–sided generalized Gaussian distribution (GGD), i.e.,

$$\begin{cases} p(x|\theta) & \sim & N(0,\theta); \\ p(\theta) & = & \frac{a\ \nu\eta(\nu)}{\Gamma(1/\nu)} e^{-[a\,\eta(\nu)\,\theta]^{\nu}} \quad \text{for} \quad \theta \geq 0, \end{cases} \tag{13}$$

where ν is the *shape parameter*, and $\eta(\nu) \stackrel{\triangle}{=} \sqrt{\frac{\Gamma(3/\nu)}{\Gamma(1/\nu)}}$. Here $\Gamma(\cdot)$ is the Gamma function. For $\nu = 1$, we have the Laplacian model in (12). For decreasing values of ν, the tails of the distribution become increasingly flat. This model allows more flexibility than (12) and

is often more realistic for practical images. See for instance Fig. 7a, which plots the log–histogram of the local variance for the finest horizontal subband of *Lena*. The local variance is estimated from the eight adjacent wavelet coefficients. The single–sided GGD with $\nu = 0.6$ (dash–dotted line) provides a good fit for the histogram. In Fig. 7b, we observe that the normalized wavelet coefficients follow a distribution very close to a Gaussian ($N(0,1)$). This is consistent with the model (13).

Given the prior $p(x,\theta)$, $I(X;\theta)$ is evaluated via numerical integration. Fig. 8 plots the value of $I(X;\theta)$ for shape parameter $\nu \in [0.3,\ 2]$ in the model (13). For the finest horizontal subband of *Lena*, with $\nu = 0.6$, $I(X;\theta) = 0.406$ bits. From Table 1, the estimate $\hat{I}(X;T) = 0.322$ bits. Hence the upper bound (10) is reasonably tight (again assuming $\hat{I}(X;T) \approx I(X;T)$).

- **Model 2**:

$$\begin{cases} p(x|\theta = \sigma^2) & \sim & N(0,\sigma^2); \\ p(\sigma) & = & ae^{-a\sigma} \ \text{ for } \sigma \geq 0. \end{cases} \tag{14}$$

This model assumes the local standard deviation σ is exponentially distributed with parameter $a > 0$. The marginal distribution $p(x)$ has a heavy tail. Fig. 9a plots the histogram of local standard deviation for the finest horizontal subband of *Barbara*. The exponential distribution fits this histogram quite closely. The conditional prior $p(x|t = \sigma^2) \sim N(0,\sigma^2)$ is quite realistic, as shown in Fig. 9b.

Under this model, the distribution of the variance $\theta = \sigma^2$ is $p(\theta) = \frac{a}{2\sqrt{\theta}} e^{-a\sqrt{\theta}}$ for $\theta \geq 0$. The scaling $\sigma' = a\sigma$ and $x' = ax$ gives a doubly stochastic model independent of a. Hence the mutual information $I(X;\theta)$ is again independent of a. Its value is obtained by numerical integration: 0.796 bits, regardless of the value of the parameter a. This upper bound is again quite close to the value $\hat{I}(X;T) = 0.696$ bits obtained using our nonparametric estimation technique (see Table 1) for the same *Barbara* subband.

- **Model 3**:

$$\begin{cases} p(x|\theta) & \sim & N(0,\theta); \\ p(\theta) & \sim & \text{Uniform } [0,a]. \end{cases} \tag{15}$$

Under this model, the marginal distribution $p(x) = \sqrt{\frac{2}{\pi a}} e^{-\frac{x^2}{2a}} - \frac{|x|}{a} + \frac{|x|}{a}\ \text{erf}(\frac{|x|}{\sqrt{2a}}))$, where $\text{erf}(\cdot)$ is the error function. For large $|x|$, this distribution tends to a Gaussian. Under this

model, $I(X;\theta) = 0.193$ bits regardless of the value of a. Compared to (12) and (14), the model (15) is less realistic. We include this model for reference: It gives the value of $I(X;\theta)$ when the wavelet coefficients follow a particular fast-decaying (not as "heavy–tailed" as the previous models) prior $p(x)$.

Under the doubly stochastic models (12), (13), (14), and (15), the optimal summarizing function $f(\cdot)$ takes the form of (6) with equal weights, as this is the sufficient statistics for estimating θ from the neighborhood $\mathcal{N}X$.

The approach above provides a useful alternative to the more complicated nonparametric mutual information estimation method described in Sec. 5. Under a given model $p(x,\theta)$, $I(X;\theta)$ provides an upper bound on the intrascale mutual information $I(X;\mathcal{N}X)$ via (10). As discussed in Sec. 4, for the bound (10) to be tight, one would need sufficiently large $\mathcal{N}X$, in which the data X and $\{\mathcal{N}X_i\}$ are independently generated from the distribution $p(\cdot|\theta)$. Then T is a good estimator of θ. Moreover, in order to have $\hat{I}(X;T) \approx I(X;T)$, we need a large number of samples from the distribution $p(\theta)$. The assumptions above are satisfied if θ viewed as a function of wavelet coefficient location is a slowly-varying, stationary, ergodic random field with the prescribed marginal distribution.

7 Discussion

In practical image processing applications, it is often not clear how to select a statistical model, from which an appropriate algorithm can be derived. In this paper, competing models are compared using information–theoretic metrics. We have evaluated mutual information for interscale, intrascale, and composite wavelet models. Mutual information may also be useful in practical image processing algorithms; see recent research on image registration [30, 31], blind source separation [32, 33], and image restoration [34] for examples. How to incorporate the mutual information analysis efficiently in a compression or estimation algorithm is not discussed in this paper, and would be an interesting subject for future research.

Mutual information provides bounds on the performance of compression, estimation, and classification algorithms. This is attractive both from a theoretical and a practical point of view because bounds provide an economical alternative to extensive runs (tests) of image processing algorithms.

Such considerations have contributed to the popularity of Fisher information and Cramer-Rao bounds in statistical signal processing. The image modeling tools developed in this paper are applicable not only to standard compression, denoising and classification problems, but also to more complicated imaging applications.

Our analysis has been developed for complete wavelet representations. An extension to overcomplete wavelet representation would follow similar principles but appears to be more difficult. The coefficients in oversampled subbands exhibit strong dependencies, thus the summarization function $T = f(\mathcal{N}X)$, which should ideally be a sufficient statistic, should take a different form. The form of T depends on the modeling of the overcomplete wavelet domain coefficients, which is not a mature subject yet.

Acknowledgment. The authors are grateful to Prof. Alfred Hero for stimulating discussions on entropy estimation and the distortion–rate bound. The authors would like to thank the anonymous reviewers for making constructive comments and suggestions.

A Numerical methods to estimate mutual information

Method 1. To estimate mutual information $I(X;Y)$, compute the histogram of (X,Y) using a uniform discretization of $\log X$ and $\log Y$. We use a log nonlinearity due to the wide spread of the wavelet coefficients. Recall that $I(X;Y) = I(f_1(X); f_2(Y))$ for any invertible functions f_1 and f_2 such as the log function.

Method 2 (Darbellay and Vajda [44]). This is an iterative method which recursively partitions the data into nonuniform cells. The approach is data–dependent and adaptive. Each partitioning operation splits a cell into four quadrants. Consider the joint density of (X,Y) in the cell (normalized to integrate to 1). The partitioning algorithm splits the cell at the median of X along the range of X, and splits at the median of Y along the range of Y. If the contribution of a quadrant to $\hat{I}(X;Y)$ falls below a prespecified threshold δ, X and Y are considered approximately independent inside the quadrant, and further splitting of the quadrant is prohibited. On the other hand, if the contribution is above δ, then the quadrant is further partitioned. All four quadrants are partitioned independently. The procedure stops when no cell is subject to further splitting.

Overall, the partitioning is nonuniform and nonseparable. After completion of the algorithm, the range of X and Y is partitioned into a collection of discretization cells. The estimate $\hat{I}(X;Y)$ is computed similarly to (11). For the estimate $\hat{I}(X;Y)$ to converge to $I(X;Y)$, the following conditions are required: $N \to \infty$; $\delta \to 0$; and the number of data samples in each discretization cell tends to infinity. For finite N, the threshold δ needs to be chosen properly — too large δ produces underestimated $I(X;Y)$, while too small δ produces overestimated $I(X;Y)$. In our experiments, we have calibrated the case where X and Y are two correlated Gaussian random variables with correlation coefficient $r = 0.1$, 0.3, 0.6, and 0.9. The threshold δ is adjusted so that the estimate of $I(X;Y)$ produced by the algorithm from N iid data samples well matches its theoretical value computed from (2). This produces $\delta = 0.005$.

References

[1] J. A. O'Sullivan, R. E. Blahut, and D. L. Snyder, "Information–theoretic image formation," *IEEE Trans. on Info. Theory*, vol. 44, pp. 2094–2123, Oct. 1998.

[2] "*IEEE Transactions on Info. Theory.*" special issue on information-theoretic imaging, Vol. 46, No. 5, Aug. 2000.

[3] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Proc.*, vol. 41, pp. 3445–3462, Dec. 1993.

[4] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 243–250, June 1996.

[5] R. L. Joshi, H. Jafarkhani, J. H. Kasner, T. R. Fischer, N. Farvardin, M. W. Marcellin, and R. H. Bamberger, "Comparison of different methods of classification in subband coding of images," *IEEE Trans. on Image Processing*, vol. 6, pp. 1473–1486, Nov. 1997.

[6] S. LoPresto, K. Ramchandran, and M. T. Orchard, "Image coding based on mixture modeling of wavelet coefficients and a fast estimation–quantization framework," in *Data Compression Conference 97*, (Snowbird, Utah), pp. 221–230, 1997.

[7] S. D. Servetto, K. Ramchandran, and M. T. Orchard, "Image coding based on a morphological representation of wavelet data," *IEEE Trans. on Image Processing*, vol. 8, pp. 1161–1174, Sept. 1999.

[8] R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Trans. on Image Processing*, vol. 8, pp. 1688–1701, Dec. 1999.

[9] E. P. Simoncelli, "Modeling the joint statistics of images in the wavelet domain," in *Proc. SPIE 44th Annual Meeting*, (Denver, CO), July 1999.

[10] M. Malfait and D. Roose, "Wavelet–based image denoising using a Markov random field *a priori* model," *IEEE Trans. on Image Processing*, vol. 6, pp. 549–565, Apr. 1997.

[11] S. G. Chang, B. Yu, and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," in *Proc. of ICIP'98*, (Chicago, IL), pp. I. 535–539, Oct. 1998.

[12] M. K. Mıhçak, I. Kozintsev, K. Ramchandran, and P. Moulin, "Low complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Signal Processing Letters*, vol. 6, no. 12, 1999.

[13] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. on Signal Processing*, vol. 46, pp. 886–902, Apr. 1998.

[14] K. E. Timmermann and R. D. Nowak, "Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging," *IEEE Trans. on Info. Theory*, vol. 45, pp. 846–862, Apr. 1999.

[15] H. Choi and R. Baraniuk, "Multiscale texture segmentation using wavelet–domain hidden Markov models," in *International Conference on Signals, Systems and Computers*, vol. 2, (Pacific Grove, CA), pp. 1692–1697, 1998.

[16] J. Li, R. M. Gray, and R. A. Olshen, "Multiresolution image classification by hierarchical modeling with two dimensional hidden Markov models," *IEEE Transactions on Info. Theory*, pp. 1826–1841, Aug. 2000.

[17] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and Bayesian restoration of images," *IEEE Trans. on PAMI*, vol. 6, pp. 721–741, Nov. 1984.

[18] C. Bouman and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Trans. on Image Processing*, vol. 2, pp. 296–310, July 1993.

[19] M. W. Marcellin, M. J. Gormish, A. Bilgin, and M. P. Boliek, "An overview of JPEG-2000," in *Data Compression Conference 2000*, (Snowbird, Utah), pp. 523–541, Mar. 2000.

[20] D. Taubman, E. Ordentlich, M. Weinberger, G. Seroussi, I. Ueno, and F. Ono, "Embedded block coding in JPEG2000," in *Proc. IEEE Int. Conf. on Image Proc.*, (Vancouver, Canada), pp. II. 33–36, Sept. 2000.

[21] J. Liu and P. Moulin, "Image denoising based on scale–space mixture modeling for wavelet coefficients," in *Proc. of ICIP'99*, (Kobe, Japan), pp. I. 386–390, Oct. 1999.

[22] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic Press, 1999.

[23] E. P. Simoncelli and E. H. Adelson, "Noise removal via Bayesian wavelet coring," in *Proc. ICIP'96*, (Lausanne, Switzerland), pp. I. 379–382, Oct. 1996.

[24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY: John Wiley and Sons, Inc., 1991.

[25] J. Ziv and M. Zakai, "On functionals satisfying a data–processing theorem," *IEEE Trans. on Info. Theory*, vol. 19, pp. 275–283, May 1973.

[26] A. O. Hero, "On the problem of granulometry for a degraded boolean image model," in *Proc. of ICIP'99*, (Kobe, Japan), pp. II. 16–20, Oct. 1999.

[27] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA: Holden-Day, 1964.

[28] V. Poor, *An Introduction to Signal Detection and Estimation, 2nd Ed.* New York, NY: Springer-Verlag, 1994.

[29] J. D. Gorman and A. O. Hero, "Lower bounds for parametric estimation with constraints," *IEEE Trans. on Info. Theory*, vol. 36, pp. 1285–1301, Nov. 1990.

[30] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," in *Proc. 5th Int. Conf. Computer Vision*, (Boston, MA), pp. 16–23, June 1995.

[31] P. Thevenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Trans. on Image Processing*, vol. 9, pp. 2089–2099, Dec. 2000.

[32] M. Girolami (ed.), *Advances in independent component analysis*. London, New York: Springer, 2000.

[33] J. Cardoso, "Blind signal separation: statistical principles," *Proc. of IEEE*, vol. 86, pp. 2009–2025, Oct. 1998.

[34] A. J. Bell and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[35] M. K. Mıhçak, T. Docimo, P. Moulin, and K. Ramchandran, "Design and analysis of a forward–adaptive wavelet image coder," in *Proc. of ICIP 2000*, (Vancouver, Canada), Oct. 2000.

[36] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice Hall, 1989.

[37] I. Daubechies, *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992.

[38] D. R. Brillinger, *Time Series: Data Analysis and Theory*. New York, NY: McGraw Hill, 1981.

[39] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, UK: Chapman and Hall, 1986.

[40] H. Stark and J. W. Woods, *Probability, Random Processes, and Estimation Theory for Engineers, second edition*. Upper Saddle River, NJ: Prentice Hall, 1994.

[41] M. K. Mıhçak, I. Kozintsev, and K. Ramchandran, "Local statistical modeling of wavelet image coefficients and its application to denoising," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, (Phoenix, AZ), pp. 3253–3256, Mar. 1999.

[42] P. Hall and S. C. Morton, "On the estimation of entropy," *Ann. Inst. Statist. Math.*, vol. 45, pp. 69–88, 1993.

[43] J. Beirlant, E. J. Dudewica, L. Gyöfi, and E. van der Meulen, "Non–parametric entropy estimation: An overview," *Int. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, 1997.

[44] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Trans. on Information Theory*, vol. 45, pp. 1315–1321, May 1999.

[45] U. Grenander, *Elements of Pattern Theory*. Baltimore: Johns Hopkins University Press, 1996.

[46] J. S. Lee, "Digital image enhancement and noise filtering by use of local statistics," *IEEE Trans. on PAMI*, vol. PAMI2, pp. 165–168, Mar. 1980.

[47] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," in *Conference on Neural Information Processing Systems (NIPS)*, (Denver, CO), pp. 855–861, Nov. 1999.

[48] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," in *Proc. of the 45th Annual Meeting of the SPIE*, (San Diego, CA), July 2000.

[49] A. Hjørungnes, J. Lervik, and T. Ramstad, "Entropy coding of composite sources modeled by infinite Gaussian mixture distributions," in *1996 IEEE Digital Signal Processing Workshop*, (Loen, Norway), pp. 235–238, Sept. 1996.

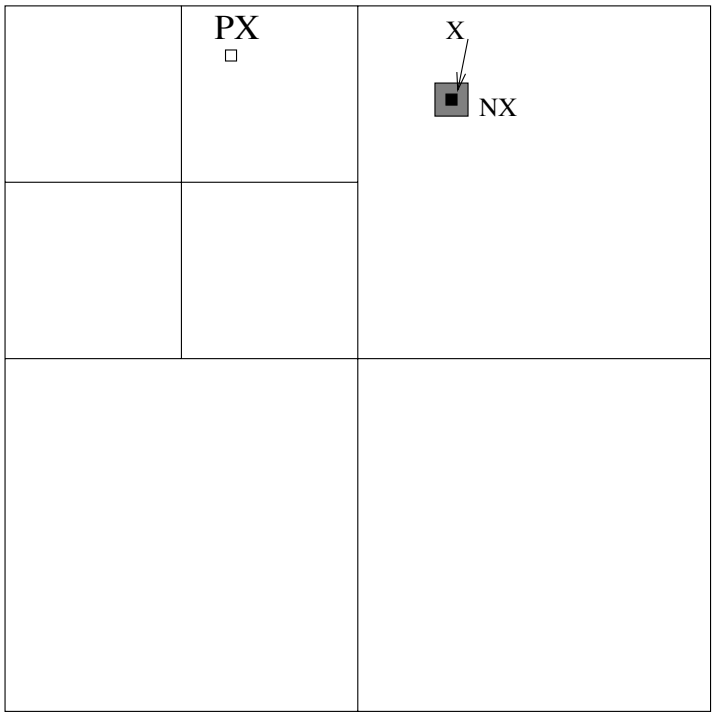


Figure 1: Definition of $\mathcal{P}X$ and $\mathcal{N}X$. For a wavelet coefficient X (pictured as the little black block), $\mathcal{P}X$ is its parent in the coarser band, and $\mathcal{N}X$ is its neighborhood, pictured as the gray area surrounding X.

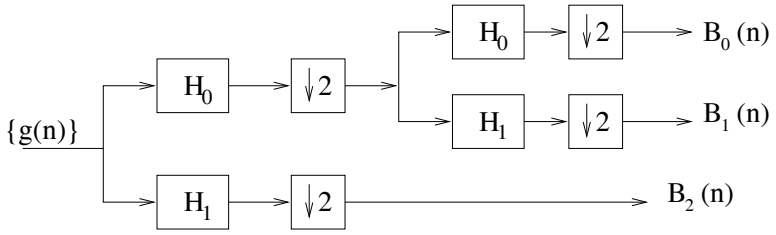
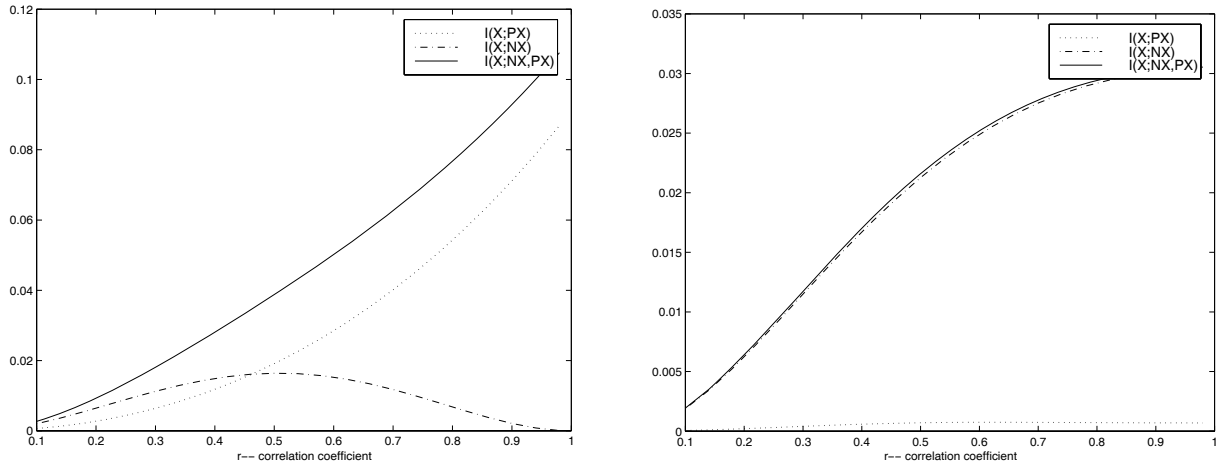
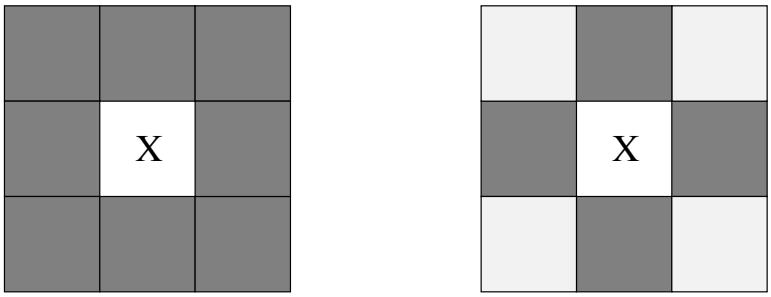


Figure 2: Wavelet decomposition of a 1–D signal, using lowpass and highpass filters $H_0(z)$ and $H_1(z)$.



(a) (b)

Figure 3: Mutual informations $I(X;\mathcal{P}X)$, $I(X;\mathcal{N}X)$, and $I(X;\mathcal{P}X,\mathcal{N}X)$ for the wavelet coefficients of a stationary AR–1 Gaussian processes, as functions of r. (a) Using Haar wavelets; (b) using Daubechies' 8–tap filters.



(a) (b)

Figure 4: Weights assigned to the neighborhood $\mathcal{N}X$. (a) Equal weights; (b) symmetric weights. Neighbors marked with the same grey level are assigned the same weights.

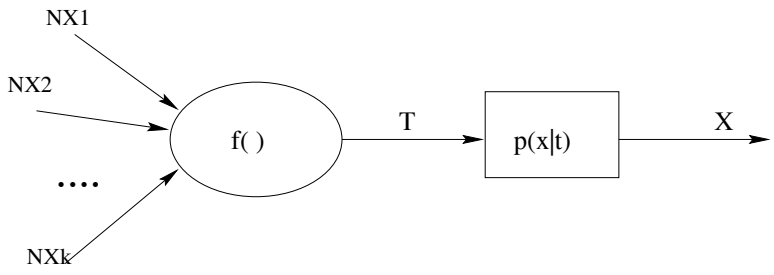


Figure 5: Reduction of dimensionality through a many–to–one mapping.

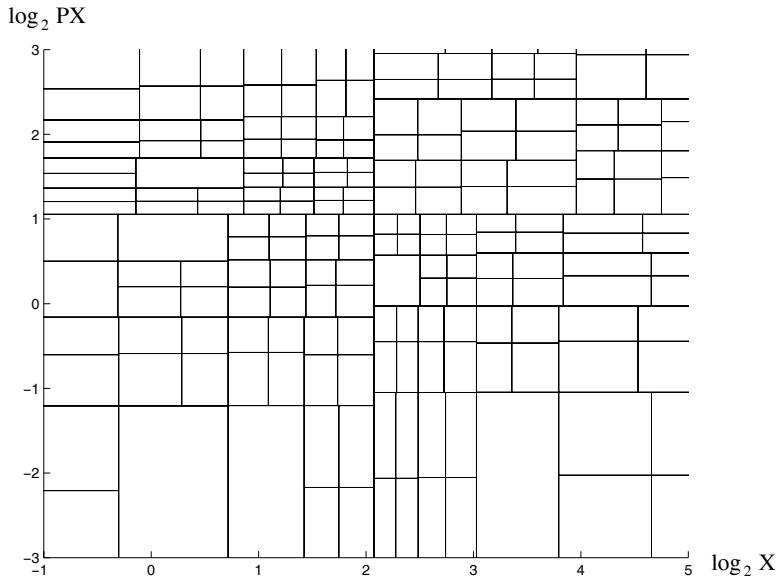
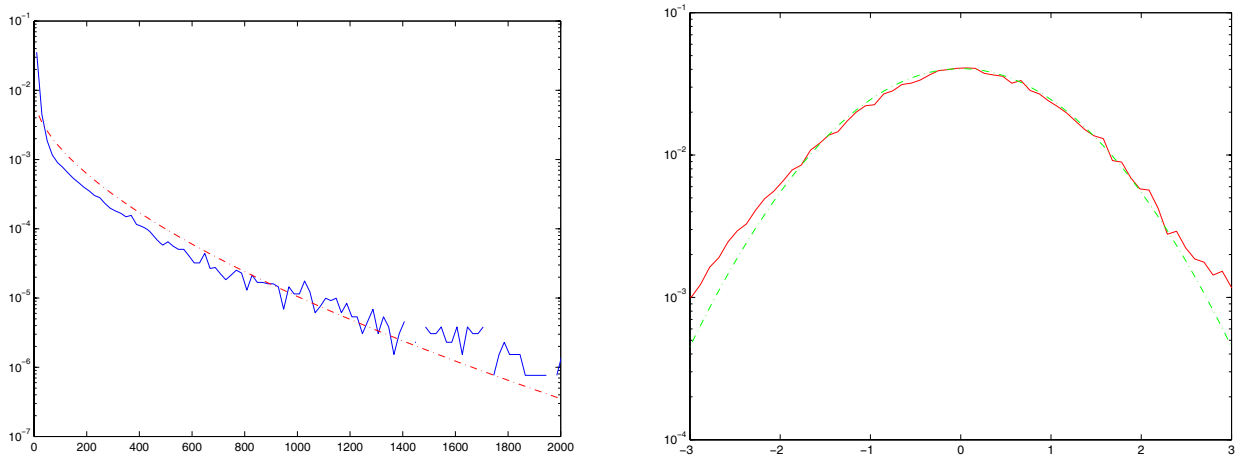


Figure 6: The partitioning result of the method by Darbellay and Vajda [44] when computing $I(X, \mathcal{P}X)$ for the finest horizontal subband of *Lena*. The horizontal axis is $\log_2 X$ for $X > 0$, and the vertical axis is $\log_2 \mathcal{P}X$ for $\mathcal{P}X > 0$.



(a) (b)

Figure 7: (a) Histogram of estimated local variance (solid line), approximated using a single–sided GGD prior (13) with $\nu = 0.6$ (dash–dotted line); (b) histogram of the normalized wavelet coefficients (solid line) approximated using an $N(0, 1)$ prior (dash–dotted line), all plotted in log scale. The histogram is for the finest horizontal subband of *Lena*.

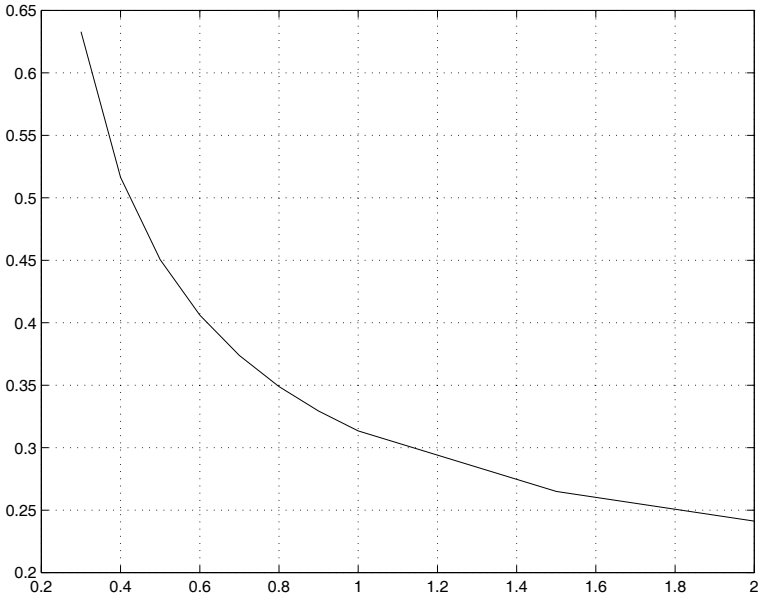
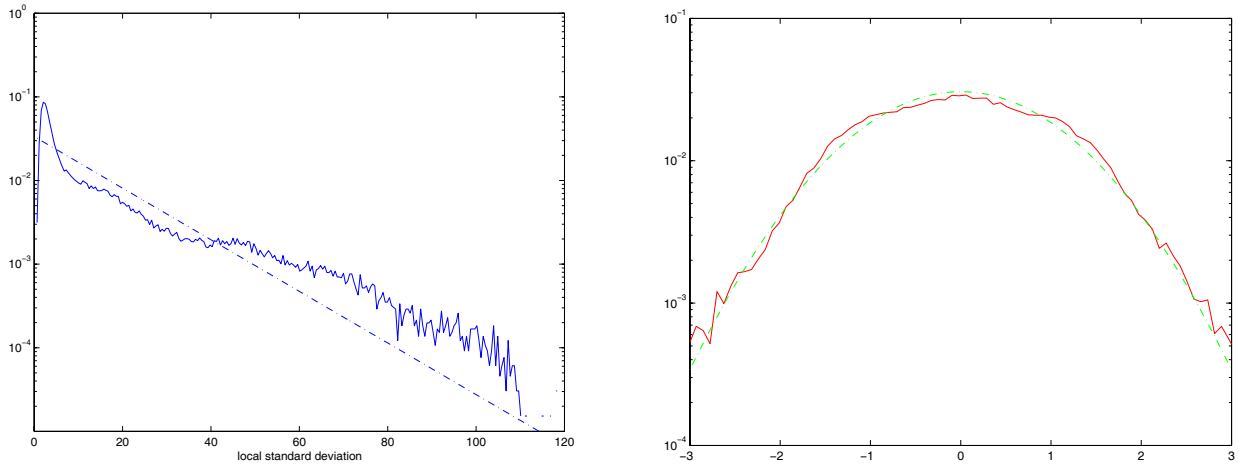


Figure 8: The mutual information $I(X;\theta)$ for the doubly stochastic model with $p(x|\theta) \sim N(0,\theta)$ and $\theta \sim$ single–sided GGD (13) with shape parameter ν. The horizontal axis is ν, varying from 0.3 to 2; the vertical axis is $I(X;\theta)$.



(a) (b)

Figure 9: (a) Histogram of estimated local standard deviation (solid line), approximated using an exponential prior (dash–dotted line); (b) histogram of the normalized wavelet coefficients (solid line) approximated using an $N(0,1)$ prior (dash–dotted line), all plotted in log scale. The histogram is for the finest horizontal subband of *Barbara*.

Lena, finest subbands			
	$\hat{I}(X;\mathcal{P}X)$	$\hat{I}(X;T)$	$\hat{I}(X;T,\mathcal{P}X)$
horizontal, $W_i = 1/8$	0.195	0.322	0.352
vertical, $W_i = 1/8$	0.144	0.239	0.264
diagonal, $W_i = 1/8$	0.084	0.135	0.159
horizontal, $\beta = 0.25$	0.195	0.330	0.359
vertical, $\beta = 0.60$	0.144	0.244	0.269
diagonal, $\beta = 0.65$	0.084	0.136	0.160
Barbara, finest subbands			
	$\hat{I}(X;\mathcal{P}X)$	$\hat{I}(X;T)$	$\hat{I}(X;T,\mathcal{P}X)$
horizontal, $W_i = 1/8$	0.206	0.696	0.723
vertical, $W_i = 1/8$	0.155	0.464	0.497
diagonal, $W_i = 1/8$	0.225	0.491	0.522
horizontal, $\beta = 0.45$	0.206	0.706	0.732
vertical, $\beta = 0.60$	0.155	0.488	0.519
diagonal, $\beta = 0.65$	0.225	0.498	0.529
Peppers, finest subbands			
	$\hat{I}(X;\mathcal{P}X)$	$\hat{I}(X;T)$	$\hat{I}(X;T,\mathcal{P}X)$
horizontal, $W_i = 1/8$	0.157	0.280	0.312
vertical, $W_i = 1/8$	0.158	0.286	0.320
diagonal, $W_i = 1/8$	0.054	0.138	0.161
horizontal, $\beta = 0.50$	0.157	0.286	0.318
vertical, $\beta = 0.35$	0.158	0.294	0.326
diagonal, $\beta = 0.35$	0.054	0.143	0.167

Table 1: Comparison of estimated mutual informations using equal weights and adaptive weights in (6). These numbers are calculated using Daubechies 4–tap filters and a log scale histogram.

Lena, next–to–finest subbands			
	$\hat{I}(X;\mathcal{P}X)$	$\hat{I}(X;T)$	$\hat{I}(X;T,\mathcal{P}X)$
horizontal, $W_i = 1/8$	0.275	0.500	0.587
vertical, $W_i = 1/8$	0.235	0.409	0.468
diagonal, $W_i = 1/8$	0.200	0.366	0.424
Barbara, next–to–finest subbands			
	$\hat{I}(X;\mathcal{P}X)$	$\hat{I}(X;T)$	$\hat{I}(X;T,\mathcal{P}X)$
horizontal, $W_i = 1/8$	0.130	0.530	0.599
vertical, $W_i = 1/8$	0.165	0.485	0.556
diagonal, $W_i = 1/8$	0.109	0.645	0.711
Peppers, finest subbands			
	$\hat{I}(X;\mathcal{P}X)$	$\hat{I}(X;T)$	$\hat{I}(X;T,\mathcal{P}X)$
horizontal, $W_i = 1/8$	0.268	0.411	0.486
vertical, $W_i = 1/8$	0.254	0.377	0.455
diagonal, $W_i = 1/8$	0.148	0.232	0.298

Table 2: Mutual information in the next–to–finest subbands. These numbers are calculated using Daubechies 4–tap filters and a log scale histogram.

	finest subbands		next–to–finest subbands	
	Improvement over interscale	Improvement over intrascale	Improvement over interscale	Improvement over intrascale
Lena	84%	13%	85%	16%
Barbara	201%	6%	383%	12%
Peppers	133%	13%	61%	22%
Baboon	209%	9%	274%	21%
Bank	67%	14%	77%	19%
Lake	157%	12%	112%	16%
Couple	102%	13%	129%	20%
Plane	86%	9%	86%	15%
Milkdrop	149%	14%	135%	19%
Tiffany	111%	10%	118%	19%

Table 3: Summary over 10 images: comparison of composite mutual information $\hat{I}(X;T,\mathcal{P}X)$ with interscale and intrascale mutual informations ($\hat{I}(X;\mathcal{P}X)$ and $\hat{I}(X;T,\mathcal{P}X)$). The second and third columns show the average improvements for the finest subbands (horizontal, vertical, and diagonal); and the last two columns show the averages over next–to–finest subbands. The second and fourth columns display the percentage by which $\hat{I}(X;T,\mathcal{P}X)$ is larger than $\hat{I}(X;\mathcal{P}X)$; the third and fifth columns display the percentage by which $\hat{I}(X;T,\mathcal{P}X)$ is larger than $\hat{I}(X;T)$.

	Lena		*Barbara*		*Peppers*	
	$\hat{I}(X;\mathcal{P}X)$	$\hat{I}(X;T)$	$\hat{I}(X;\mathcal{P}X)$	$\hat{I}(X;T)$	$\hat{I}(X;\mathcal{P}X)$	$\hat{I}(X;T)$
log scale histogram	0.195	0.322	0.206	0.696	0.157	0.280
adaptive partitioning [44]	0.181	0.290	0.208	0.632	0.136	0.238

Table 4: Comparison of two methods to estimate the mutual informations: the log scale histogram method and the adaptive partitioning method [44]. The mutual informations are for the finest horizontal subband. Equal weights are used, and $\mathcal{N}X$ contains the eight coefficients adjacent to X.

	$\hat{I}(X;\mathcal{P}X)$	$\hat{I}(X;T)$	$\hat{I}(X;T,\mathcal{P}X)$
horizontal, $W_i = 1/3$	0.195	0.256	0.318
vertical, $W_i = 1/3$	0.144	0.179	0.241
diagonal, $W_i = 1/3$	0.084	0.090	0.137

Table 5: Results with causal neighborhood $\mathcal{N}X$ defined as the left, upper, and upper–left neighbors of X. The mutual informations are for the finest subbands of *Lena*. Equal weights are used in (6).