Published in: **IEEE Transactions on Image Processing**, 8(12):1688--1701, Dec 1999.
Originally: GRASP Laboratory Technical Report #414, University of Pennsylvania.
Written: 30 May 1997. Revised: 14 Oct 1998. Final revision: 2 May 1999.
Available at: ftp://ftp.cis.upenn.edu/pub/eero/buccigrossi97.ps.gz

Image Compression via Joint Statistical Characterization in the Wavelet Domain

Robert W. Buccigrossi
GRASP Laboratory
Dept. of Computer & Information Science
University of Pennsylvania
Philadelphia, PA 19104
butch@grip.cis.upenn.edu

Eero P. Simoncelli
Center for Neural Science, and
Courant Inst. of Mathematical Sciences
New York University
New York, NY 10003
eero.simoncelli@nyu.edu

Abstract

We develop a probability model for natural images, based on empirical observation of their statistics in the wavelet transform domain. Pairs of wavelet coefficients, corresponding to basis functions at adjacent spatial locations, orientations, and scales, are found to be non-Gaussian in both their marginal and joint statistical properties. Specifically, their marginals are heavy-tailed, and although they are typically decorrelated, their magnitudes are highly correlated. We propose a Markov model that explains these dependencies using a linear predictor for magnitude coupled with both multiplicative and additive uncertainties, and show that it accounts for the statistics of a wide variety of images including photographic images, graphical images, and medical images. In order to directly demonstrate the power of the model, we construct an image coder called EPWIC (Embedded Predictive Wavelet Image Coder), in which subband coefficients are encoded one bitplane at a time using a non-adaptive arithmetic encoder that utilizes conditional probabilities calculated from the model. Bitplanes are ordered using a greedy algorithm that considers the MSE reduction per encoded bit. The decoder uses the statistical model to predict coefficient values based on the bits it has received. Despite the simplicity of the model, the rate-distortion performance of the coder is roughly comparable to the best image coders in the literature.

Many applications in image processing require a prior probability model. This is especially true for the application of image compression, in which the theoretical limits of an algorithm are determined by the underlying prior model. In this paper, we describe an explicit prior probability model for photographic images, and test this model by using it as the basis for an image compression algorithm. The resulting algorithm is quite flexible, and well-suited for encoding of images that must be retrieved over a variety of communication links.

Wavelet representations, in which images are decomposed using basis functions localized in spatial position, orientation, and spatial frequency (scale), have proven to be extremely effective for image compression [e.g., 33, 35, 8, 1, 25, 23]. We believe there are several statistical reasons for this success.

• RB was supported by NSF Graduate Fellowship GER93-55018 and the GRASP Laboratory at the University of Pennsylvania. EPS was supported by NSF CAREER grant MIP-9796040, ARO/MURI DAAH04-96-1-0007, and the Sloan Center for Theoretical Neurobiology at NYU.
• Preliminary versions of this work have been published in [3] and [31].

The most widely known of these is that wavelet transforms are reasonable approximations to the Karhunen-Loève expansion for fractal signals [36], such as natural images [22]. The subbands of an orthonormal wavelet decomposition have a wide range of variances whose sum is equal to that of the original image. If the subbands are encoded with a simple first-order entropy encoder, the minimum coding size of the image representation is sum of the entropies of the subbands. Since entropy is a concave function, the differences in subband variances result in a coding cost significantly less than the first-order entropy of the original image pixels.

In addition to this redistribution of variance, the coefficients of wavelet transforms have significantly non-Gaussian marginal statistics for typical images, and thus have lower entropy than a Gaussian-distributed signal of the same variance. This property has been exploited in compression, noise removal and texture synthesis [e.g., 15, 6, 10, 9, 38, 30]. We discuss it in greater detail in section 1, and provide an explicit model for these marginals.

Finally, wavelet decompositions exhibit joint statistical regularities that have been utilized in a number of recent image coding algorithms [17, 25, 21, 24, 12, 4, 37]. These regularities are the primary topic of this paper. We discuss them in greater detail in section 2, and develop an explicit model to describe the relationships between coefficients of different subbands.

In order to demonstrate the quality of our statistical model, the latter half of the paper describes an embedded predictive wavelet image coder that directly utilizes the model. Section 3 describes the compression algorithm, and the details of the coder implementation. Finally, section 4 analyzes the performance of the coder, and compares it to several standard coders.

1 First-order Subband Statistics

A number of authors have observed that wavelet subband coefficients have highly non-Gaussian statistics [e.g., 15, 7, 16, 30]. Histograms[1] for subbands of separable wavelet decompositions of several images are plotted in figure 1. Compared to a Gaussian, these densities are more sharply peaked at zero, with more extensive tails. The intuitive explanation for this is that images typically have spatial structure consisting of smooth areas interspersed with occasional edges or other abrupt transitions. The smooth regions lead to near-zero coefficients, and the structures give occasional large-amplitude coefficients.

To quantify this, we give the sample kurtosis κ (fourth moment divided by squared second moment) below each histogram. The estimated kurtoses of all of the subbands are significantly larger than the value of three expected for a Gaussian distribution. These examples were computed for subbands of an orthonormal separable wavelet decomposition (see section 3 for details), but we find that they are similar for any octave-bandwidth subbands.

These non-Gaussian densities should be contrasted with statistics of frequency-based decompositions which are approximately Gaussian. Since the Gaussian is the maximal-entropy distribution for a given variance, wavelet-based coders are able to achieve higher degrees of compression than frequency-based coders such as JPEG. The non-Gaussianity of wavelet marginals may be taken as an indication that the Wavelet basis is more appropriate for image representation than either pixel or Fourier repre-

[1] By considering these as representative of the underlying coefficient densities, we are making implicit assumptions of strict-sense stationarity and ergodicity.

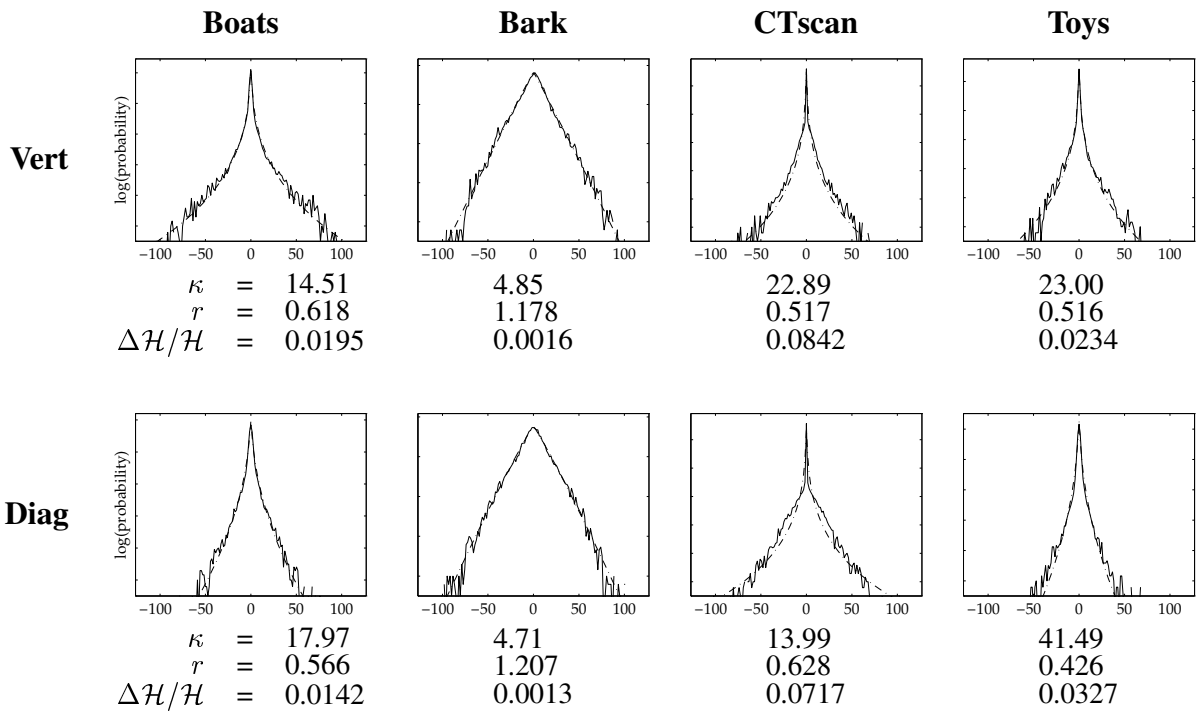

Figure 1: Examples of 256-bin subband coefficient histograms (solid lines) fitted with the density of equation (1) (dotted lines), plotted in the log domain. Subbands correspond to four different images (a landscape, a texture, a medical image, and a synthetic image) and two different orientations (top: vertical, bottom: diagonal), at the second highest frequency scale. Below each graph is the sample kurtosis, κ (fourth moment divided by squared variance), the model exponent, r, and the relative entropy, $\Delta\mathcal{H}$ (Kullback-Leibler divergence) between the histogram and the model as a fraction of the empirical (histogram) entropy, $\mathcal{H}$. The Bark image had the most Gaussian marginals in our image set, and the CTscan image gave the worst model fit (in terms of relative entropy).

sentations.

Wavelet coefficient marginals have been previously modeled [15, 10, 30, 13, 11] using a two-parameter "generalized Laplacian" (or "stretched exponential") density function of the form:[2]

$$f_{s,r}(c) = \frac{e^{-|c/s|^r}}{N(s,r)}, \tag{1}$$

where $N(s,r) = 2s\Gamma(1/r)/r$, and $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$, the Gamma function. The parameters $\{s,r\}$ are directly related to the second and fourth moments. Specifically:

$$\sigma^2 = \frac{s^2\Gamma(\frac{3}{r})}{\Gamma(\frac{1}{r})}, \qquad \kappa = \frac{\Gamma(\frac{1}{r})\Gamma(\frac{5}{r})}{\Gamma^2(\frac{3}{r})}, \tag{2}$$

where σ^2 is the distribution variance, and κ is the kurtosis.

For each subband, we solve (numerically) for the parameters $\{s,r\}$ by minimizing the relative entropy (i.e., the Kullback-Leibler divergence) between a discretized model distribution and the 256-bin

[2]This model is appropriate for the bandpass coefficient marginals. The lowpass subband coefficients are modeled using a uniform distribution.

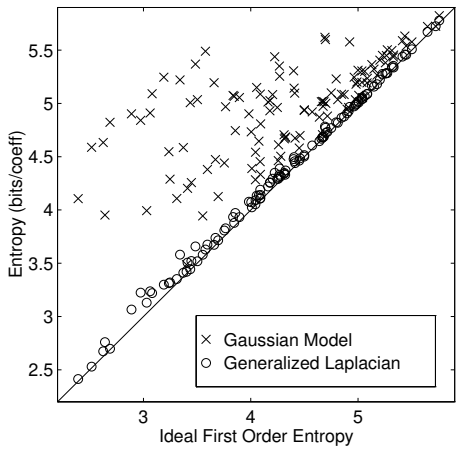

Figure 2: Comparison of encoding costs. Plotted are encoding cost assuming the generalized Laplacian density of equation (1) (O's), and the encoding cost assuming a Gaussian density (X's), versus the encoding cost using a 256-bin histogram. Points are plotted for 9 bands (3 scales, 3 orientations) of the 13 images in the sample set of figure 10. The average relative entropy (Kullback-Leibler divergence) of the Gaussian model is 0.592 bits/coefficient, while the average relative entropy of the generalized Laplacian model is 0.035 bits/coefficient.

coefficient histogram:

$$\Delta\mathcal{H}(s,r) = -\sum_{n=1}^{256} h_n \log_2 \frac{\bar{f}_{s,r}(c_n)}{h_n},$$

where $\bar{f}_{s,r}(c_n)$ is the integral of the density given in equation (1) over the nth histogram bin (centered at value c_n), and h_n is the normalized histogram count (frequency) for the nth histogram bin. The measure $\Delta\mathcal{H}(s,r)$ corresponds to the cost (in bits) of encoding the data with an entropy coder that assumes the distribution $f_{s,r}(c)$. For the images in our sample set, σ^2 is roughly proportional to β^l (where l indicates the scale or "pyramid level"), with $\beta \in [3.5, 10]$. The exponent r is typically in the range $[0.5, 1.0]$, corresponding to kurtosis values in the range $[6, 25.2]$.

We make no claim of optimality for this model: Other authors [e.g., 38] have used alternative density functions to describe these distributions. Nevertheless, the fits are surprisingly good. Figure 1 shows the log-domain plots of the sample histograms together with plots of the fitted density function of equation (1). We have included both the best and worst cases from the set of images in our test set (shown in figure 10). Below each figure is the relative entropy between the histograms and fitted densities.

Figure 2 shows a scatterplot comparing the encoding cost using the model of equation (1), and the encoding cost assuming a Gaussian density vs. the encoding cost assuming accurate knowledge of a 256-bin histogram. The Gaussian examples were computed with the distribution variance matched to the sample variance. Note that the relative entropy of the generalized Laplacian model is less than 0.25 bits/coefficient for our sample images, as compared with the Gaussian density model which often has a relative entropy greater than 1.0 bit/coefficient.

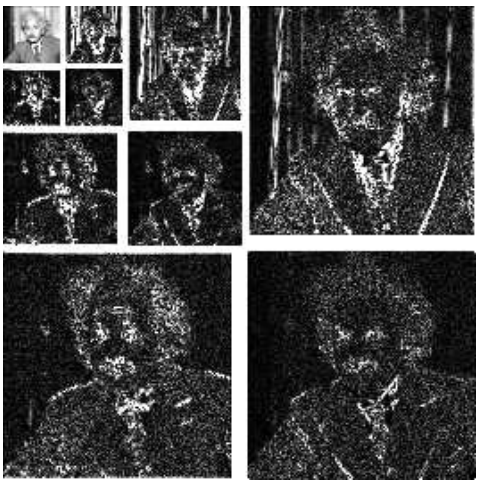

Figure 3: Coefficient magnitudes of a wavelet decomposition. Shown are absolute values of subband coefficients at three scales, and three orientations of a separable wavelet decomposition of the Einstein image. Also shown is the lowpass residual subband (upper left). Note that high-magnitude coefficients of the subbands tend to be located in the same (relative) spatial positions.

2 Joint Subband Statistics

As mentioned in the introduction, the coefficients of wavelet subbands are approximately decorrelated. Nevertheless, it is clear from visual inspection that wavelet coefficients are *not* statistically independent. Figure 3 shows the magnitudes of wavelet coefficients in a four-scale decomposition. Large-magnitude coefficients tend to occur at neighboring spatial locations, and also at the same relative spatial locations of subbands at adjacent scales and orientations [3].

Spatial and scale-to-scale dependencies are utilized implicitly in a number of recent image compression schemes. Shapiro [25] constructed the Embedded Zerotree Wavelet (EZW) coder to exploit the fact that a coefficient is likely to have small magnitude if the coefficients at coarser scales have small magnitudes. The Zerotree technique encodes entire trees of zeros with a single symbol, thus capturing a portion of the conditional distribution of a coefficient given its coarser scale neighbors (parent, grandparent, etc). Several authors [18, 17, 21] have used vectorized lookup tables to predict blocks of fine coefficients from blocks of coarse coefficients. Schwartz et. al. [24] used adaptive entropy coding to capture conditional statistics of coefficients based on the most significant bits of each of the eight spatial neighbors and the coefficient at a coarser scale. Chrysafis and Ortega [4] switch between multiple probability models depending on values of neighboring coefficients. Said and Pearlman [23] use a predictive scheme to give high-quality zerotree coding results, and Wu and Chen [37] have extended the EZW coder to use local coefficient "contexts". LoPresto et. al. [13] model the coefficients as being chosen from a generalized Laplacian density and estimate the model parameters from local neighborhoods. Joshi et. al. [11] adaptively condition the encoding of classification maps of regions of coefficients based upon the classes of the left and parent regions.

2.1 Joint Magnitude Statistics

We wish to explicitly examine and utilize the statistical relationship between wavelet coefficient magnitudes. Consider two coefficients representing information at adjacent scales, but the same orientation (e.g, horizontal) and spatial location. As in the previous section, we will assume strict-sense stationarity and ergodicity, which allows us to consider the joint histogram of this pair of coefficients, gathered over the spatial extent of the image, as representative of the underlying statistics. Figure 4A shows the conditional histogram $h\left(C|P\right)$ of the "child" coefficient conditioned on the coarser-scale "parent" coefficient. The histogram illustrates several important aspects of the relationship between the two coefficients. First, they are (approximately) second-order decorrelated, since the expected value of C is roughly zero for all values of P. Second, the variance of C exhibits a strong dependence on the value of P. Thus, although C and P are uncorrelated, *they are still statistically dependent*. Furthermore, this dependency cannot be eliminated through further linear transformation.

The structure of the relationship between C and P becomes more apparent upon transforming to the log domain. Figure 4B shows the conditional histogram $h\left(\log_2(C')|\log_2(Q)\right)$, where $Q = |P|$ and $C' = |C|$. The right side of the distribution is unimodal and concentrated along a unit-slope line. This suggests that in this region, the conditional expectation, $\mathcal{E}(C'|Q)$, is approximately proportional to Q. Furthermore, vertical cross sections (i.e., conditional histogram for a fixed value of Q) have approximately the same shape for different values of Q. Finally, the left side of the distribution is concentrated about a horizontal line, suggesting that C' is independent of Q in this region. We suspect these low-amplitude coefficients are dominated by quantization errors and other sources of uncertainty.

The intuition for the right side of the distribution is that typical localized image structures (e.g., edges) tend to have substantial power across many scales at the same spatial location. These structures will be represented in the wavelet domain via a superposition of basis functions at these scales. The signs and relative magnitudes of the coefficients associated with these basis functions will depend on the precise location, orientation and scale of the structure. The absolute magnitudes will also scale with the contrast of the structure. Thus, measurement of a large coefficient at one scale means that large coefficients at adjacent scales are more likely.

The form of the histograms shown in figure 4 is surprisingly robust across a wide range of images. Furthermore, the qualitative form of these statistical relationships also holds for pairs of coefficients at adjacent spatial locations (which we call "siblings"), adjacent orientations ("cousins"), and adjacent orientations at a coarser scale ("aunts"). This set of potential conditioning coefficients (we refer to these as "neighbors") is illustrated in figure 5.

2.2 Linear Magnitude Predictor

Given the linear relationship between the magnitudes of large-amplitude coefficients, and the difficulty of characterizing the full multi-dimensional density, we chose to examine a linear predictor for coefficient magnitude:

$$l(\vec{Q}) \equiv \vec{w} \cdot \vec{Q} = \sum_k w_k Q_k, \tag{3}$$

where the coefficient magnitude set $\{Q_k\}$ corresponds to a subset of the potential conditioning neighbors, as depicted in figure 5. For a single subband, the weights w_k used to compute the predictor are

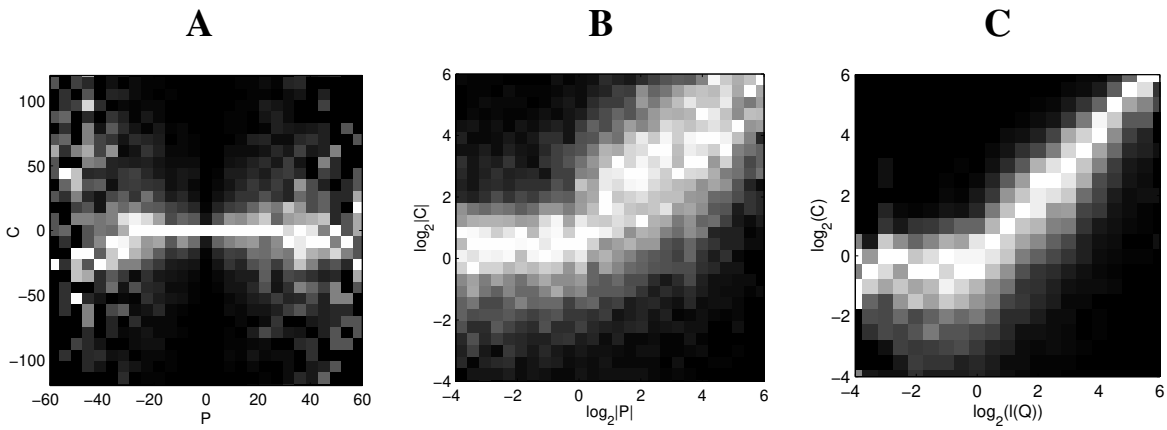

Figure 4: Conditional histograms for a fine scale horizontal coefficient from the Boats image. Brightness corresponds to probability, except that each column has been independently rescaled to fill the full range of display intensities. **A:** Conditioned on the Parent (same location and orientation, coarser scale) coefficient. **B:** Same as **A**, but in the log domain. **C:** Conditioned on a linear combination of neighboring coefficient magnitudes.

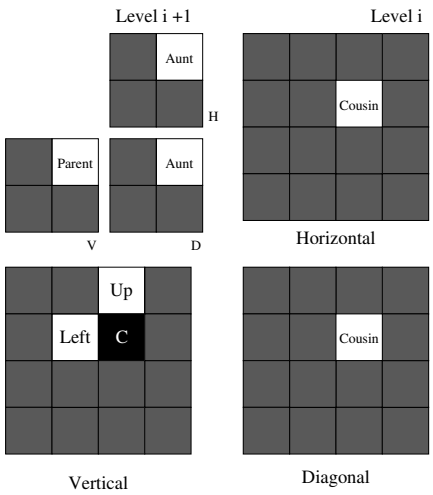

Figure 5: Subset of wavelet coefficients surrounding a given coefficient (C) that are potentially suitable for conditioning.

Child subband	Last neighbor included in predictor					
Horizontal	Left	Up	Parent	DiagCousin	LeftLeft	DiagAunt
	0.3322	0.4308	0.4635	0.4804	0.4903	0.4939
Vertical	Up	Left	Parent	DiagCousin	UpUp	DiagAunt
	0.3513	0.4356	0.4675	0.4865	0.4929	0.4987
Diagonal	Up	Left	Parent	Horiz Cousin	Vert Cousin	Left Left
	0.2175	0.2792	0.3134	0.3235	0.3294	0.3356

Table 1: Cumulative mutual information between coefficient magnitude, C, and a linear combination of neighbor magnitudes, $l(\vec{Q})$. Each entry gives the mutual information for a subset containing the neighbors indicated at the top of that column and all columns to the left. Notice that the local neighbors within the subband (Left and Up), the Parent, and the Cousins contribute most to the mutual information. Values are averaged over the two finest pyramid scales of three training images (Lena, Boats, Baboon).

chosen to minimize the expected squared error. That is:

$$\vec{w} = \mathcal{E}(\vec{Q}\vec{Q}^T)^{-1} \cdot \mathcal{E}(C' \cdot \vec{Q}), \tag{4}$$

where $\mathcal{E}(\cdot)$ indicates the expected value of a random variable, C' corresponds to the coefficient magnitude being estimated, and $\vec{Q}$ is a vector containing the magnitudes of the conditioning neighbors. In practice, the expectation is estimated by summing spatially over the subband.

Figure 4C shows a conditional histogram, $h\left(\log_2(C')|\log_2(l(\vec{Q}))\right)$ based on magnitudes of eight adjacent coefficients in the same subband, two cousin coefficients, and one parent coefficient (interpolated to the correct position using bilinear interpolation). Note that the distribution has a similar appearance to the single-parent distribution of figure 4A, but the linear region is extended and the conditional variance is greatly reduced.

In order to determine which coefficients to include in the conditioning set, we calculated the mutual information between C and $l(\vec{Q})$ for a variety of choices of interband and intraband coefficients in $\{Q_k\}$. The mutual information gives the theoretical coding gain (in bits per coefficient) obtained when encoding C using the conditional histogram $h\left(C \mid l(\vec{Q})\right)$ (i.e., assuming $l(\vec{Q})$ is known to the receiver) compared with encoding C using only the marginal histogram $h(C)$. Rather than exhaustively explore all possible neighbor subsets, we used a greedy algorithm. Specifically, the set is constructed incrementally: at each step, we incorporate the remaining neighbor whose inclusion maximizes the mutual information. Table 1 shows the greedy optimal neighbor subset for the three oriented subbands. Using this analysis, and imposing causality (assuming a standard scanline ordering of the coefficients), we decided to include neighbors corresponding to the first four table columns when coding the horizontal and vertical bands, and the first five columns for the diagonal bands.

2.3 Conditional Probability Model

We wish to construct a probability model for a coefficient conditioned on its neighbors. We were surprised to observe that the conditional distribution in the log domain, when normalized for mean and variance, is highly consistent across subbands of an image, and even across a wide range of

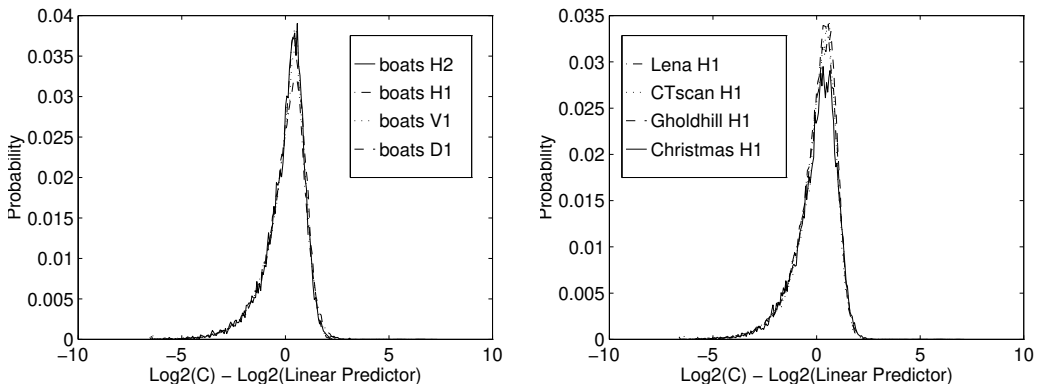

Figure 6: Comparison of conditional distributions in the log domain of different subbands and images. Distributions were normalized (in the log domain) to have mean 0 and variance 1. **Left:** Comparison of distributions for different subbands of the Boats image. **Right:** Comparison of distributions for different images (Lena, Goldhill, CTscan, Christmas).

images. Figure 6 shows a comparison of these conditional distributions for four subbands from the "Boats" image, and also a comparison of a single band across four different images. We included only the right portion of each conditional histogram (i.e., the region in which C' is proportional to $l(\vec{Q})$).

The fact that the conditional histograms seem to have a constant shape that shifts linearly with the predictor in the log domain suggests a model of multiplicative uncertainty. In particular, we use the following model for the conditional density:

$$C = M \cdot l(\vec{Q}) + N, \tag{5}$$

where $l(\vec{Q})$ is the linear magnitude predictor described previously, and M and N are two mutually independent zero-mean random variables. This implies that the variance of C is a linear function of $l^2(\vec{Q})$. In addition to the Markov assumption, we also assume that the coefficient is conditionally independent of the neighbors, given the value of the linear predictor: $\mathcal{P}\left(C \mid \{Q_n\}\right) = \mathcal{P}\left(C \mid l(\vec{Q})\right)$.

To model the distribution of M, we used a discretized lookup table obtained by averaging the mean- and variance-normalized conditional histograms (as shown in figure 4) of three training images (Lena, Boats, Baboon), at two scales (levels 2 and 3) and all three orientations. We assume N is independent of M, and Gaussian-distributed. Given these distributional assumptions, the model described by equation (5) is characterized by the linear weights $\{w_k\}$, and the variance, σ^2, of N.

In order to fit the model to a given set of data, the linear weights are chosen via equation (4) to be least-squares optimal. The variance, σ^2 is then estimated by minimizing the relative entropy between the joint model density and the joint histogram. Figure 7 shows comparisons of joint histograms of the second-level horizontal subband of four different images, with plots of the fitted density function generated by equation (5). The estimated densities are a reasonable fit, although several of the actual histograms show a narrowing of the conditional density for large predictor values. We believe this is due to small amounts of residual linear correlation between coefficients.

An entropy calculation shows the value and quality of the model. Figure 8 shows a scatterplot comparing encoding cost based on the joint probability model of equation (5) vs. the encoding cost assuming

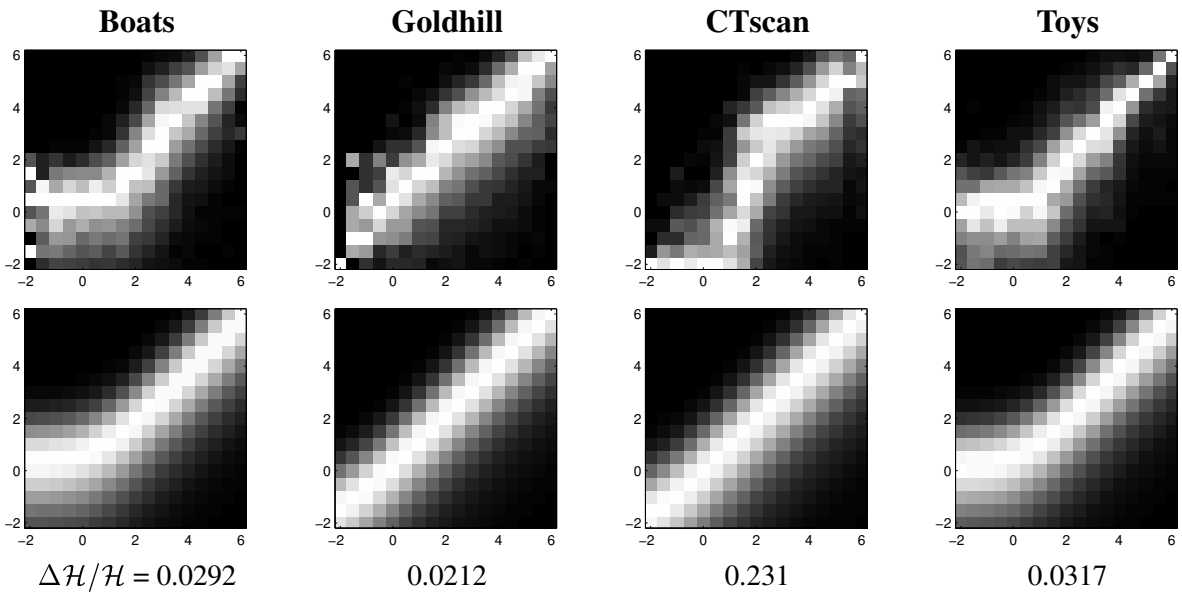

Figure 7: Top: Examples of log-domain conditional histograms for the second-level horizontal subband of different images, conditioned on an optimal linear combination of coefficient magnitudes from adjacent spatial positions, orientations, and scales. **Bottom**: Model of equation (5) fitted to the conditional histograms in the top row. Intensity corresponds to probability, except that each column has been independently rescaled to fill the full range of intensities. Also given is the relative entropy, $\Delta\mathcal{H}$, between the histogram and the model as a fraction of the histogram entropy, $\mathcal{H}$.

precise knowledge of a 256×256-bin histogram. Also included is a comparison to the first-order histogram entropies. The conditional model falls short of the empirical entropy by less than 0.7 bit. In these situations, the empirical conditional histogram for large-magnitude predictors is sparse and has high variance. The predictive models, however, are based on *smooth* high-variance densities. Thus, the empirical values are deceptively low due to detailed knowledge of the coefficient values for this specific subband. Nevertheless, the linear-predictive model is substantially better than the first-order model, consistent with the mutual entropy estimates of table 1.

Finally, we should consider the signs of the coefficients. As mentioned in the previous section, we model the lowpass coefficient distribution as a uniform density. The values are almost entirely positive: Over our sample set of 13 images, only 2.4% of the coefficients are negative. For the bandpass subbands, the probability of positive and negative coefficients is equal. They are, however, not spatially independent. In the horizontal bands, for example, the probability of the "Up" neighbor having the same sign is 36%. We utilize this simple single-neighbor conditioning in our coder. There are also more complex relationships between sign bits in neighboring subbands, but we do not attempt to characterize those in this paper.

3 Implementation of a Progressive Image Coder

In this section we describe the implementation of our Embedded Predictive Wavelet Image Coder (EPWIC), based on the conditional probability model developed in Section 2.3. Our implementation

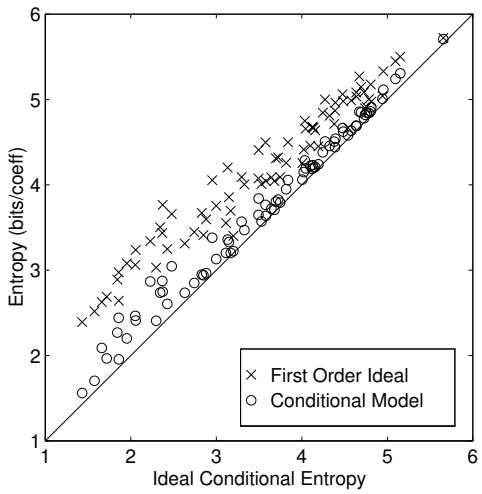

Figure 8: Comparison of encoding cost using the conditional probability model of equation (5), and the encoding cost using the first-order histogram as a function of the encoding cost using a 256×256-bin joint histogram. Points are plotted for 6 bands (2 scales, 3 orientations) of the 13 images in our sample set. The average relative entropy (Kullback-Leibler divergence) of the empirical marginal histogram is 0.548 bits/coefficient, while the average relative entropy of the conditional model is 0.129 bits/coefficient.

is simple and reasonably efficient, but comes quite close to the theoretical entropy associated with the probability model. In addition, its performance is roughly comparable to the current best coders in the image processing literature.

3.1 Separable Wavelet Decomposition

We utilize a recursive pyramid decomposition based on separable 9/7 tap biorthogonal filter set of [2] which satisfies the one-dimensional system diagram shown in figure 9. These filters have become quite popular in the compression literature. The two one-dimensional kernels, $L_9(\omega)$ and $H_7(\omega)$, are applied separably along the axes of the image sampling lattice in order to generate a single level of a wavelet pyramid. This consists of lowpass, vertical, horizontal and diagonal subbands. Subsequent pyramid levels (i.e., subbands at different scales) are created by applying this 4-band splitting procedure recursively to the lowpass subband. Convolution boundaries are handled by symmetric reflection of the image about the edge pixels, as described in [27]. Reconstruction is achieved as shown in the diagram, using filters related to the analysis filters via the expressions:

$$\begin{aligned} H_9(\omega) &= e^{j\omega} L_9(\omega + \pi) \\ H_7(\omega) &= e^{j\omega} L_7(\omega + \pi) \end{aligned}$$

We denote the basis functions in the separable wavelet transform as $w_s(x-n, y-m)$, where s indicates the subband (determined by the orientation and scale) and (n, m) indicates the spatial location of the basis function. All functions are scaled to have unity L_2-norm. The wavelet representation consists of the set of coefficients, $\{c_s(n, m)\}$, associated with these basis functions.

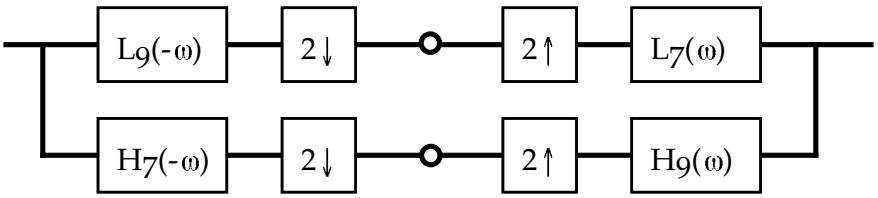

Figure 9: One-dimensional analysis/synthesis diagram for the dyadic biorthogonal wavelet decomposition used in EPWIC.

3.2 Coefficient Bitplane Encoding

In order to have maximal control over the ordering of image information, we map each subband coefficient to a 16-bit binary integer (including sign bit). That is:

$$c_s(n,m) = a_s \rho_s(n,m) \sum_k 2^k b_s(n,m,k),$$

where a_s is a scalar multiplier for subband s (used to rescale the values for 16-bit integer representation), $\rho_s(n,m)$ is the sign (± 1), and $b_s(n,m,k)$ corresponds to the kth bit of the coefficient $c_s(n,m)$.

The wavelet decomposition describes the image $I(x,y)$ as a linear combination of the basis functions:

$$\begin{aligned} I(x,y) &= \sum_{s,n,m} a_s \ \rho_s(n,m) \ \sum_k 2^k \ b_s(n,m,k) \ w_s(x-n, y-m) \\ &= \sum_{s,n,m,k} b_s(n,m,k) \ \left[a_s \rho_s(n,m) 2^k w_s(x-n, y-m) \right]. \end{aligned}$$

The second expression suggests that we may view this as a type of representation in which the coefficients are restricted to the set $\{0,1\}$ (i.e., they are single-bit quantities). The basis functions of this representation are $w'_s(n,m,k) = a_s \rho_s(n,m) 2^k w_s(x-n, y-m)$, which are related to each other by translation, dilation, Fourier modulation, negation, and *multiplication by powers of two*.

Progressive transmission of an image requires us to choose an ordering of the coefficient bits. In order to keep the complexity of the coder down, we assume that all bits at a given significance level k of a subband will be sent consecutively, in raster order. We refer to this collection of bits as a *bitplane*. We also assume that the bitplanes of a given subband will be sent in order from most to least significant. Since most of the coefficient values are close to zero, the sign bit of each coefficient is sent only when needed, immediately after the first non-zero magnitude bit, as in [24].

In general, the ordering of bitplanes across subbands should take into account both the encoded size of the bitplane and the improvement in decoded image quality resulting from the incorporation of that bitplane. We use a greedy algorithm (which we refer to as a "bang-for-the-buck" algorithm), in which we select at each step the remaining bitplane that gives the maximum reduction in mean squared error (MSE) per encoded bit. That is, we choose the bitplane that produces the steepest descent of the rate-distortion curve. Wang and Kuo [34] use a similar technique designed specifically for successive approximation quantization, in which they related the steepest rate-distortion curve with the highest quantization threshold. Our "bang-for-the-buck" algorithm is a generalization of this concept.

The bitplanes are encoded with a static arithmetic encoder whose probabilities are determined directly from our image model. The algorithm is similar to that described in [20], which encodes a data stream using a probability distribution that is adaptively computed and stored in a histogram. Instead

of computing such a histogram, our encoder uses the distribution specified by our statistical model. Since the "symbols" of our input stream are single bits, the probability that a bit is non-zero is all that is needed to construct the arithmetic code.

3.3 Calculation of Bit Probabilities

Our encoding technique makes direct use of the model joint probability density described earlier. In particular, both encoder and decoder must use this distribution to compute the conditional mean estimate for each coefficient, given the bits that have been sent/received thus far. In addition, as described above, the arithmetic coder and decoder must know the probability that a given bit will be non-zero.

Consider the arithmetic encoding of the kth bit of a particular subband coefficient, C_k. The encoder must calculate the probability that the bit is nonzero, given the set of all coefficient bits that have already been received. The set of bits received constrains the magnitude of the coefficient of interest, C', and constraints the magnitudes of each of the conditioning coefficients, $\{Q_n\}$, to lie in particular ranges:

$$\begin{aligned} C' &\in [l_0, h_0] \\ Q_n &\in [l_n, h_n], \qquad n = 1, 2, \ldots, N. \end{aligned} \tag{6}$$

The probability that we wish to calculate is:

$$\begin{aligned} &\mathcal{P}\left(C_k = 1 \,|\, \text{bits received thus far}\right) \\ &\quad = \mathcal{P}\left(m_0 < C' < h_0 \,|\, l_0 < C' < h_0, l_n < Q_n < h_n,\ \forall n\right) \\ &\quad = \frac{\mathcal{P}\left(m_0 < C' < h_0 \,|\, l_n < Q_n < h_n,\ \forall n\right)}{\mathcal{P}\left(l_0 < C' < h_0 \,|\, l_n < Q_n < h_n,\ \forall n\right)} \\ &\quad = \frac{\int dp\, \mathcal{P}\left(m_0 < C' < h_0 \,|\, l(\vec{Q}) = p\right)\ \mathcal{P}\left(l(\vec{Q}) = p \,|\, l_n < Q_n < h_n,\ \forall n\right)}{\int dp\, \mathcal{P}\left(l_0 < C' < h_0 \,|\, l(\vec{Q}) = p\right)\ \mathcal{P}\left(l(\vec{Q}) = p \,|\, l_n < Q_n < h_n,\ \forall n\right)}, \end{aligned} \tag{7}$$

where $m_0 = (l_0 + h_0)/2$. The last expression uses the assumption of our joint probability model: given $l(\vec{Q})$, the individual components of $\vec{Q}$ do not provide additional information about C'. Thus,

$$\mathcal{P}\left(l_0 < C' < h_0 \,|\, l(\vec{Q}) = p, l_n < Q_n < h_n,\ \forall n\right) \approx \mathcal{P}\left(l_0 < C' < h_0 \,|\, l(\vec{Q}) = p\right).$$

In order to avoid the computationally expensive integration over p, we introduce two approximations in the conditional probability of C' that allow us to perform a simple one-dimensional calculation. First, we assume that the density of C' (see equation (5)) has a constant shape, independent of the magnitude predictor $l(\vec{Q})$. In this case, we may divide C' by its standard deviation, $\sqrt{l^2(\vec{Q}) + \sigma^2}$, to get a scalar random variable that is independent of $l(\vec{Q})$. We compute a lookup table, $G(\cdot)$, containing the average of the mean- and variance-normalized log-domain cumulative histograms of this quantity for three training images (Lena, Boats, Baboon) at two scales (levels 2 and 3) and all three orientations. We can then use $G(\cdot)$ as a parameterized model for the conditional cumulative:

$$\mathcal{P}\left(C' < c \,|\, l(\vec{Q}) = p\right) \approx G\left(\frac{\log_2(c/\sqrt{p^2 + \sigma^2})}{\alpha}\right), \tag{8}$$

where σ^2 is the variance of N, and α^2 is the second moment of $\log_2(M)$.

Second, we eliminate the need for integration by replacing the integration variable p by its conditional mean. Specifically, we rewrite the model of equation (5) in terms of the current estimate of the magnitude predictor ($\hat{p}$), and the error of that estimate (e_p):

$$C = M \cdot (\hat{p} + e_p) + N.$$

Assuming e_p is independent of M and N, the variance of this expression is $\hat{p}^2 + \sigma_{\hat{p}}^2 + \sigma^2$, where $\sigma_{\hat{p}}^2$ is the variance of the error between p and $\hat{p}$, and the numerator integral becomes:

$$\begin{aligned} &\int dp\, \mathcal{P}\left(m_0 < C' < h_0 | l(\vec{Q}) = p\right) W(p) \\ &\quad \approx \left[G\left(\frac{\log_2(h_0) - \log_2(\sqrt{\hat{p}^2 + \sigma_{\hat{p}}^2 + \sigma^2})}{\alpha}\right) - G\left(\frac{\log_2(m_0) - \log_2(\sqrt{\hat{p}^2 + \sigma_{\hat{p}}^2 + \sigma^2})}{\alpha}\right)\right] \int dp\, W(p), \end{aligned} \tag{9}$$

where

$$W(p) = \mathcal{P}\left(l(\vec{Q}) = p \,|\, l_n < Q_n < h_n,\ \forall n\right).$$

The current estimate of the predictor $\hat{p}$ is computed from the current estimates of the neighbor magnitudes $\hat{q}_n$ (described below in equation (11)):

$$\hat{p} = \sum_n w_n \hat{q}_n,$$

and the estimate error $\sigma_{\hat{p}}^2$ is defined by the magnitude error estimates $\sigma_{\hat{q}_n}^2$ (described below in equation (12)), assuming the neighbors are uncorrelated:

$$\sigma_{\hat{p}}^2 = \sum_n w_n^2 \sigma_{\hat{q}_n}^2$$

Finally, substituting the approximation of equation (9) into equation (7) and eliminating common factors gives:

$$\begin{aligned} &\mathcal{P}\left(C_k = 1 | \text{bits received thus far}\right) \\ &\quad \approx \frac{G\left(\frac{\log_2(h_0) - \log_2(\sqrt{\hat{p}^2 + \sigma_{\hat{p}}^2 + \sigma^2})}{\alpha}\right) - G\left(\frac{\log_2(m_0) - \log_2(\sqrt{\hat{p}^2 + \sigma_{\hat{p}}^2 + \sigma^2})}{\alpha}\right)}{G\left(\frac{\log_2(h_0) - \log_2(\sqrt{\hat{p}^2 + \sigma_{\hat{p}}^2 + \sigma^2})}{\alpha}\right) - G\left(\frac{\log_2(l_0) - \log_2(\sqrt{\hat{p}^2 + \sigma_{\hat{p}}^2 + \sigma^2})}{\alpha}\right)}. \end{aligned} \tag{10}$$

This is the expression used for the calculation of bit probabilities in the coder.

After a bit is sent or received for a coefficient magnitude, the conditional mean estimate $\hat{c}$ given its new range (l, h) is calculated from the joint probability density. As before, we avoid computationally expensive integration over p through the use of its estimate $\hat{p}$:

$$\hat{c} = \int_l^h dc\, \mathcal{P}\left(C' = c \,|\, l(\vec{Q}) = \hat{p}\right)\, c. \tag{11}$$

Similarly, the variance of the error between $\hat{c}$ and the actual magnitude is:

$$\sigma_{\hat{c}}{}^2 = \int_l^h dc\, \mathcal{P}\left(C' = c \,|\, l(\vec{Q}) = \hat{p}\right)\, (\hat{c} - c)^2. \tag{12}$$

In the coder, the integrals in equations (11) and (12) are approximated by summing over a uniform partitioning of the range (l, h) and calculating the probability of each bin using the cumulative density function given in equation (8).

3.4 Summary and Complexity Analysis of EPWIC

The following is an summary of the EPWIC compression algorithm.

1. Choose a termination MSE level (we typically use the variance associated with the quantization of the original image pixels).
2. Calculate the coefficients $c_s(n, m)$ of the wavelet decomposition.
3. For each subband in the decomposition, quantize coefficients (to 16 bits) and retain the quantization binsize, $1/a_s$.
4. Characterize the statistics for each subband:
   (a) Calculate least-squares optimal weights $\vec{w}$ of the linear predictor $l(\vec{Q})$ using equation (4).
   (b) Calculate σ^2 and α^2 that minimize the relative entropy between the joint coefficient histogram and the probability density described by equation (5).
5. Transmit an EPWIC identification tag (16 bits), the number of levels (scales) in the pyramid (3 bits), width and height of the image (5 bits each), and σ^2 (8 bits, representing $[2^{-5}, 2^5]$).
6. While the decoded MSE is greater than the termination MSE:
   (a) Determine which of the set of candidate bitplanes (i.e., the most significant remaining bitplane of each subband) should be encoded next, by comparing the "bang-for-the-buck" (MSE reduction / encoding size). For each candidate bitplane (typically one per subband):
      - Compute the conditional means of all coefficients in the subband assuming the bitplane is sent (using equation (11)). Compute the reduction in MSE that results when the image is reconstructed from the resulting wavelet pyramid.
      - For each bit in the bitplane, calculate the probability of the bit being nonzero (using equation (10)), and construct a code stream using the arithmetic coder.
   (b) Update the conditional variances of the subband to be transmitted (using equation (12)).
   (c) Transmit a tag identifying the subband to which the bitplane belongs ($\log_2$(number of subbands) bits).
   (d) If this is the first encoded bitplane of this subband, transmit:
      - The quantization binsize $1/a_s$ (7 bits, representing the interval $[2^{-15}, 2^{16}]$),
      - $\{w_k\}$ (8 bits each, representing $[-0.1, 1.1]$).
      - and α^2 (8 bits, representing $[2.7, 4]$).

(e) Transmit the encoded data.

The memory overhead of the algorithm is quite reasonable. The Wavelet pyramid requires as much space as the original image, as do the current estimates of coefficient conditional mean and variance. In addition, a lookup table containing 153 floating point numbers is used for the conditional cumulative function, $G(\cdot)$.

For our implementation of EPWIC running in MatLab on a 300 megahertz Pentium workstation, encoding a 512×512 image to 64Kbytes using a 5-level pyramid takes approximately 5 minutes. Roughly 7 seconds of this time is used to estimate the parameters, 16 seconds for arithmetic encoding of the bitplanes, and 4.5 minutes are used to calculate the "bang-for-the-buck" for the next candidate bitplanes. In order to accommodate the testing of non-orthonormal filters, the MSE was calculated by reconstructing the pyramid twice, with and without the candidate bitplane. Thus, the calculation cost could be significantly reduced through use of orthonormal filters, for which the MSE improvement can be estimated directly from the coefficients. Decoding the image takes approximately 16 seconds, 8 of which are used in the calculation of the magnitude estimates and the conditional probabilities.

4 Results

In order to demonstrate the performance of EPWIC, we encoded the set of 13 images shown in figure 10. Each image was decomposed into a discrete wavelet pyramid containing 5 scales. For comparison purposes, we considered four other image coders:

1. EPWIC-1: we implemented a progressive encoder utilizing the marginal (generalized Laplacian) density of equation (1) as a model of the first order distribution. The coder is otherwise similar to the conditional implementation (EPWIC-2), in that it uses the same greedy algorithm for ordering of bitplanes, and uses the same arithmetic coding scheme.
2. JPEG: we used version 5b of CJPEG, a standard non-progressive JPEG image coder from the Independent JPEG Group.
3. EZW[3]: as described in [25].
4. SPIHT: as described in [23].

Table 2 lists the PSNR values for EPWIC-2, EPWIC-1, EZW, SPIHT, and JPEG for five of the images. It should be noted that these PSNR values were calculated directly from the decompressed images, and the bitrates indicate actual encoding sizes, not entropy estimates. We were surprised to find that EPWIC-1 surpasses EZW for most images, since the model for this coder incorporates no joint statistical information, while EZW exploits some of the joint conditional relationships between coefficients at different scales. EPWIC-2 surpasses EZW at nearly all compression levels, and approaches the encoding capability of SPIHT at the higher compression rates. Figure 11 summarizes these results, by showing the PSNR of each coder (relative to EZW), averaged over the 13 images in our set. EPWIC-1 outperforms EZW for most compression ratios by about 0.3dB, and EPWIC-2 outperforms EZW by 0.5dB at 1 Kbyte, and nearly 1.5dB at 16 Kbytes and above.

[3] We thank the David Sarnoff Research Center for their assistance in the EZW comparisons.

Figure 10: Full set of grayscale images used in our experiments. Left to right, top to bottom: Baboon, Bark, Boats, Brain, Brownie, Cantrell, Earth, Flowers, Goldhill, Lena, MtWill, Vein, and Wedding. Images contain 512×512 8-bit pixels.

		Bits/Pixel								
Image	**Coder**	0.008	0.016	0.031	0.063	0.125	0.25	0.5	1.0	2.0
Boats	EPWIC-2	21.67	23.36	25.21	27.24	29.72	32.97	37.04	41.69	46.96
	SPIHT	22.28	23.75	25.46	27.46	29.85	33.07	37.16	41.73	46.81
	EPWIC-1	21.55	23.30	24.98	26.92	29.09	31.95	35.80	39.97	45.11
	EZW	21.34	22.83	24.81	26.86	28.87	31.69	35.58	40.00	45.82
	JPEG	NA	NA	18.29	21.83	27.76	30.88	34.63	39.10	43.54
Baboon	EPWIC-2	18.89	19.48	19.89	20.64	21.61	23.19	25.25	28.86	34.49
	SPIHT	19.18	19.57	19.98	20.74	21.72	23.27	25.65	29.17	34.98
	EPWIC-1	18.94	19.47	19.92	20.55	21.46	22.95	24.90	28.25	33.69
	EZW	18.89	19.37	19.87	20.36	21.50	22.53	24.90	28.45	34.06
	JPEG	NA	NA	16.51	19.01	20.87	22.03	24.13	26.65	30.97
Lena	EPWIC-2	21.68	23.54	25.70	28.03	30.85	33.78	37.15	40.34	45.06
	SPIHT	22.07	23.95	25.95	28.36	31.10	34.12	37.23	40.43	45.11
	EPWIC-1	21.26	23.35	25.41	27.61	30.18	33.04	35.99	39.27	44.04
	EZW	21.03	22.99	25.01	27.46	30.26	33.32	36.46	39.79	44.64
	JPEG	NA	NA	17.95	21.92	28.24	31.42	34.84	37.95	41.62
Goldhill	EPWIC-2	22.00	23.54	24.93	26.59	28.23	30.08	32.83	36.25	41.78
	SPIHT	22.56	23.89	25.26	26.70	28.39	30.45	32.99	36.44	41.99
	EPWIC-1	22.12	23.63	24.95	26.52	28.07	29.92	32.38	35.73	40.98
	EZW	21.66	23.41	24.66	26.10	27.88	29.73	31.92	35.09	39.82
	JPEG	NA	NA	17.92	23.96	26.85	29.18	31.59	34.46	38.46

Table 2: PSNR values $10 \cdot \log_{10}(255^2/\text{MSE})$ at different compression ratios for EPWIC-2, EPWIC-1, EZW, SPIHT, and JPEG. Original images are shown in figure 10.

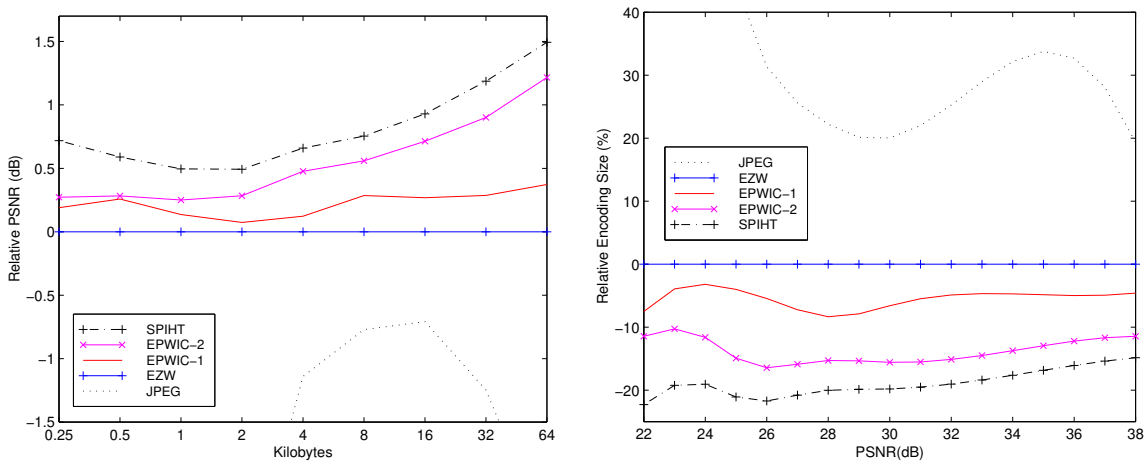

Figure 11: Relative rate-distortion tradeoff for five image coders (JPEG, EZW, SPIHT, EPWIC-1, and EPWIC-2). **Left**: PSNR values (in dB), relative to EZW (horizontal line), as a function of the number of encoded bytes. **Right**: Number of bytes necessary to achieve a given PSNR, relative to EZW (horizontal line). All curves are averages over the set of 13 images shown in figure 10.

Also shown in figure 11 is the encoding size (relative to that of EZW) as a function of target SNR. This gives a sense of how long one would wait during progressive transmission for a result of a given quality. For example, EZW would have a transmission time roughly 15% higher than EPWIC-2 for an image quality of 26dB.

In figure 12, an EPWIC-2 progressive transmission series is given for the Boats image. Wavelet aliasing artifacts are quite noticeable in the early stages of the transmission: these are a consequence of using a critically sampled subband representation. At 16 Kbytes (compared with an original image size of 256 Kbytes), the reconstructed image is remarkably close to the original. We find that EPWIC-2 compressed images are visually indistinguishable from SPIHT compressed images at all bitrates.

One would like to know how much encoding performance is being lost in the integration approximations of equation (10), and how much is lost in the overhead of sending the model parameters. In order to measure the cost of the approximations, we developed a non-progressive version of the coder called "EPWIC-2 NP". Instead of encoding the pyramid bitplane by bitplane, EPWIC-2 NP encodes the subbands simultaneously, at a fixed quantization level. In order to measure the cost of the model parameters, the PSNR values for EPWIC-2 NP were re-calculated without the overhead of encoding the parameters.

Figure 13 shows the results of this analysis. The non-progressive coder gives an improvement of roughly 0.1 dB over progressive EPWIC-2. Thus, we conclude that the integration approximations do not greatly penalize the progressive encoding performance. In addition, removing the overhead of sending the model parameters improves the performance at lower encoding sizes. In particular, EPWIC-2 NP is comparable to SPIHT for encoding sizes up to 8K.

Figure 12: Progressive decoding of the Boats image. Each image is an approximation of the original image computed by decoding the indicated number of bytes from an EPWIC-2 code stream.

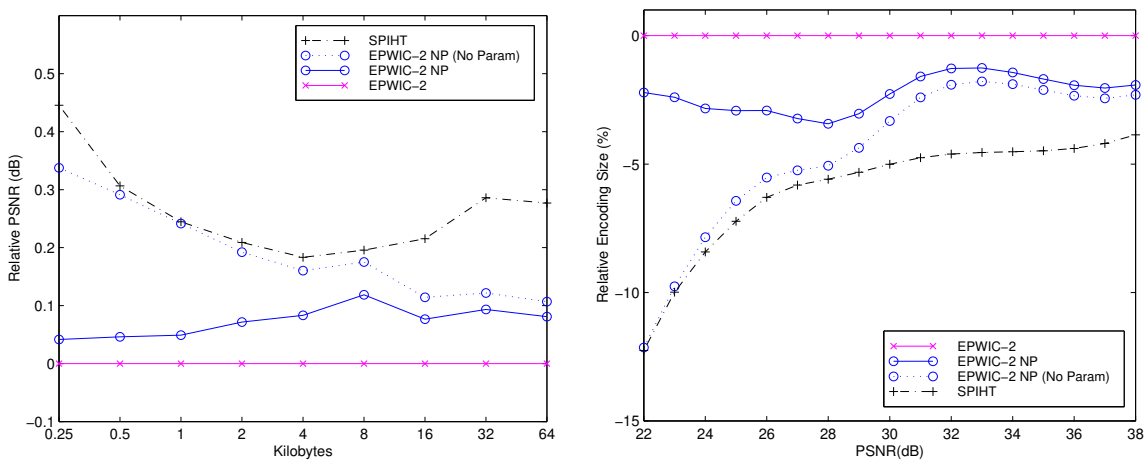

Figure 13: Coding loss due to EPWIC-2 integration approximations and parameter overhead. EPWIC-2 NP is a non-progressive version of EPWIC-2, which avoids the integration approximations of equation (9). Also shown are the PSNR values for EPWIC-2 NP, excluding the overhead of encoding the model parameters. **Left**: PSNR values (in dB), relative to EPWIC-2 (horizontal line), as a function of the number of encoded bytes. **Right**: Number of bytes necessary to achieve a given PSNR, relative to EPWIC-2 (horizontal line). All curves are averages over the set of 13 images shown in figure 10.

5 Conclusion

We have presented a conditional probability model for images based on a linear combination of the magnitudes of neighboring coefficients in a wavelet decomposition. The model characterizes the magnitude statistics of a wide variety of images, and provides a useful framework for understanding the compression capabilities of other coders. We have demonstrated the power of the model by using it explicitly in the implementation of an image coder (EPWIC-2). The compression results are surprisingly good, especially given the simplicity of the encoding scheme.

We believe that there two are main reasons that EPWIC-2 falls slightly below the encoding capabilities of SPHIT. First, by using a non-adaptive model of the statistics, EPWIC-2 has the overhead of encoding model parameters in the bit stream. As shown in figure 13, removal of this overhead nearly eliminates the gap in performance for low bitrates. Second, EPWIC-2 utilizes a conditional density that depends only on a single value (the linear combination of neighbor magnitudes). SPHIT and related recent encoders [e.g., 23, 37, 13, 11] utilize zero trees and adaptive conditional techniques, allowing them to take advantage of multi-dimensional joint statistical relationships.

There are a number of improvements that could be made in the implementation of EPWIC. The L_1-norm combination of neighboring magnitudes (described in equation (3)) could be replaced with an L_p-norm predictor, with p chosen to optimize the coding gain. Our preliminary examination of this possibility suggests that the resulting improvements are minimal. In addition, the overhead associated with the model parameters could be reduced by entropy coding these values. Finally, more sophisticated exploitation of sign statistics could yield significant improvements in compression. In particular, the current coder does not make predictions of coefficients before receiving the sign bits. A model that allowed prediction of sign bits from causal neighbors (including those at coarser scales), would allow the coder to fabricate image detail early in a progressive transmission sequence. This type of

prediction would also allow the creation of synthetic images with statistics matched to a given sample image.

We believe the explicit conditional probability model used in EPWIC-2 is well-suited for other image processing problems such as image denoising or enhancement, texture segmentation, and texture synthesis. We have begun to explore some of these applications, and the results are encouraging [28, 26, 29, 19]. We do find, however, that most of these applications require a translation-invariant representation, such as an overcomplete multi-scale pyramid [e.g., 32] or an image-specific adaptive basis [e.g., 14, 5].

Acknowledgments

We thank R. Bajcsy and the members of the GRASP lab for their support during the course of this work. We also thank the anonymous reviewers of this article for helpful suggestions.

References

[1] E. H. Adelson, E. P. Simoncelli, and R. Hingorani. Orthogonal pyramid transforms for image coding. In *Proc. SPIE*, volume 845, pages 50–58, Cambridge, MA, October 1987.

[2] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies. Image coding using wavelet transform. *IEEE Trans. Image Proc*, 1:105–220, April 1992.

[3] R. W. Buccigrossi and E. P. Simoncelli. Progressive wavelet image coding based on a conditional probability model. In *ICASSP*, volume IV, pages 2957–2960, Munich, Germany, April 1997. IEEE Sig Proc Society.

[4] C. Chrysafis and A. Ortega. Efficient context-based lossy wavelet image coding. In *Proc. of Data Compression Conference*, Snowbird, UT, March 1997.

[5] R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. Info. Theory*, IT-38:713–718, March 1992.

[6] D. L. Donoho. Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. In I. Daubechies, editor, *Proc Symp Appl Math*, volume 47, pages 173–205, Providence, RI, 1993.

[7] D. J. Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.

[8] H. Gharavi and A. Tabatabai. Sub-band coding of digital images using two-dimensional quadrature mirror filtering. In *Proc. of SPIE*, volume 707, pages 51–61, 1986.

[9] D. Heeger and J. Bergen. Pyramid-based texture analysis/synthesis. In *Proc. ACM SIGGRAPH*, August 1995.

[10] R. L. Joshi and T. R. Fischer. Image subband coding using arithmetic and trellis coded quantization. *IEEE Trans. Circuits and Systems for Video Technology*, 5(6):515–523, 1995.

[11] 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. Image Proc*, 6(11):1473–1486, 1997.

[12] F. Kossentini, W. Chung, and M. Smith. A jointly optimized subband coder. *IEEE Trans Image Processing*, 5(9):1311–1323, September 1996.

[13] S. M. LoPresto, K. Ramchandran, and M. T. Orchard. Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization frameqork. In *Data Compression Conference*, March 1997.

[14] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Proc.*, 41(12):3397–3415, December 1993.

[15] S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. PAMI*, 11:674–693, July 1989.

[16] B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7:333–339, 1996.

[17] A. Pentland and B. Horowitz. A practical approach to fractal-based image compression. In A. B. Watson, editor, *Digital Images and Human Vision*. MIT Press, 1993.

[18] A. P. Pentland, E. P. Simoncelli, and T. Stephenson. Fractal-based image compression and interpolation, 1992. U.S. Patent Number 5,148,497. Filed 14 Feb 1990, issued 15 Sep 1992.

[19] J. Portilla and E. Simoncelli. Texture modeling and synthesis using joint statistics of complex wavelet coefficients. In *IEEE Workshop on Statistical and Computational Theories of Vision*, Fort Collins, CO, 1999. Available at http://www.cns.nyu.edu/∼eero/publications.html

[20] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C - Second Edition*. Cambridge University Press, Cambridge, MA, 1992.

[21] R. Rinaldo and G. Calvagno. Image coding by block prediction of multiresolution subimages. *IEEE Trans Image Processing*, 4(7):909–920, July 1995.

[22] D. L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Phys. Rev. Letters*, 73(6):814–817, 1994.

[23] A. Said and W. A. Pearlman. An image multiresolution representation for lossless and lossy compression. *IEEE Trans. Image Proc*, 5(9), September 1996.

[24] E. Schwartz, A. Zandi, and M. Boliek. Implementation of compression with reversible embedded wavelets. In *Proc SPIE*, 1995.

[25] J. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans Signal Processing*, 41(12):3445–3462, December 1993.

[26] E. Simoncelli and J. Portilla. Texture characterization via joint statistics of wavelet coefficient magnitudes. In *Fifth IEEE Int'l Conf on Image Proc*, volume I, Chicago, October 4-7 1998. IEEE Computer Society.

[27] E. P. Simoncelli. Orthogonal sub-band image transforms. Master's thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, MA, May 1988. Also available as MIT Media Laboratory Vision and Modeling Technical Report #100.

[28] E. P. Simoncelli. Statistical models for images: Compression, restoration and synthesis. In *31st Asilomar Conf on Signals, Systems and Computers*, pages 673–678, Pacific Grove, CA, November 1997. IEEE Computer Society.

[29] E. P. Simoncelli. Bayesian denoising of visual images in the wavelet domain. In P. Müller and B. Vidakovic, editors, *Bayesian Inference in Wavelet Based Models*, chapter 18, pages 291–308. Springer-Verlag, New York, June 1999. Lecture Notes in Statistics, vol. 141.

[30] E. P. Simoncelli and E. H. Adelson. Noise removal via Bayesian wavelet coring. In *Third Int'l Conf on Image Processing*, pages 379–383, Lausanne, Switzerland, September 1996.

[31] E. P. Simoncelli and R. W. Buccigrossi. Embedded wavelet image compression based on a joint probability model. In *Fourth Int'l Conf on Image Proc*, volume I, pages 640–643, Santa Barbara, October 1997. IEEE Sig Proc Society.

[32] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multi-scale transforms. *IEEE Trans Information Theory*, 38(2):587–607, March 1992. Special Issue on Wavelets.

[33] M. Vetterli. Multi-dimensional sub-band coding: Some theory and algorithms. *Signal Processing*, 6(2):97–112, February 1984.

[34] H. Wang and C. J. Kuo. A multi-threshold wavelet coder (MTWC) for high fidelity image compression. In *IEEE ICIP*, pages 652–655, 1997.

[35] J. W. Woods and S. D. O'Neil. Subband coding of images. *IEEE Trans. Acoust. Speech Signal Proc.*, ASSP-34(5):1278–1288, October 1986.

[36] G. Wornell. Wavelet-based representations for the $1/f$ family of fractal processes. *Proc. IEEE*, 810(10):1428–1450, October 1993.

[37] X. Wu and J. Chen. Context modeling and entropy coding of wavelet coefficients for image compression. In *ICASSP*, Munich, April 1997.

[38] S. Zhu, Y. Wu, and D. Mumford. Filters, random fields and maximum entropy (FRAME) – towards the unified theory for texture modeling. In *IEEE Conf Computer Vision and Pattern Recognition*, June 1996.