Presented at: 31st Asilomar Conference on Signals Systems, and Computers
Pacific Grove, CA. November 2-5, 1997.

Statistical Models for Images: Compression, Restoration and Synthesis

Eero P. Simoncelli
Center for Neural Science, and
Courant Institute of Mathematical Sciences
New York University

We present a parametric statistical model for visual images in the wavelet transform domain. We characterize the joint densities of coefficient magnitudes at adjacent spatial locations, adjacent orientations, and adjacent spatial scales. The model accounts for the statistics of a wide variety of visual images. As a demonstration of this, we've used the model to design a progressive image encoder with state-of-the-art rate-distortion performance. We also show promising examples of image restoration and texture synthesis.

Vision is arguably our most important sensory system, judging from both the ubiquity of visual forms of communication, and the large proportion of the human brain devoted to visual processing. Nevertheless, it has proven difficult to establish a good mathematical definition (in the form of a statistical model) for visual images. Many applications in image processing can benefit from a statistical prior model. In addition, such a model is a necessary component of ecological theories of vision, which hypothesize that biological visual systems are constructed (through some combination of evolution and development) to decompose and represent the statistics of the visual world.

In this paper, we examine the problem of decomposing digitized images, through linear and/or nonlinear transformations, into statistically independent components. The classical approach to such a problem is Principal Components Analysis (PCA), also known as the Karhunen-Loève (KL) or Hotelling transform. This is a linear transform that removes second-order dependencies between input pixels. The most well-known description of image statistics is that their power spectra take the form of a power law [e.g., 20, 11, 24]. Coupled with a constraint of translation-invariance, this suggests that the Fourier transform is an appropriate PCA representation. Fourier and related representations are widely used in image processing applications.

EPS is currently supported by NSF CAREER grant MIP-9796040, and the Sloan Center for Theoretical Neurobiology at NYU.

1. Wavelet Marginal Statistical Model

Recently, multi-scale linear transforms such as wavelets have become popular for image representation. Typically, the basis functions of these representations are localized in spatial position, orientation, and spatial frequency (scale). For many applications, these appear to be superior to Fourier representations. We believe that there are a number of reasons for this. As with the Fourier basis, orthonormal wavelets are good at decorrelating second-order pixel statistics. In addition, image wavelet coefficients have significantly non-Gaussian marginal statistics [e.g., 11, 17], while Fourier coefficient marginals are much closer to Gaussian (and thus have higher entropy).

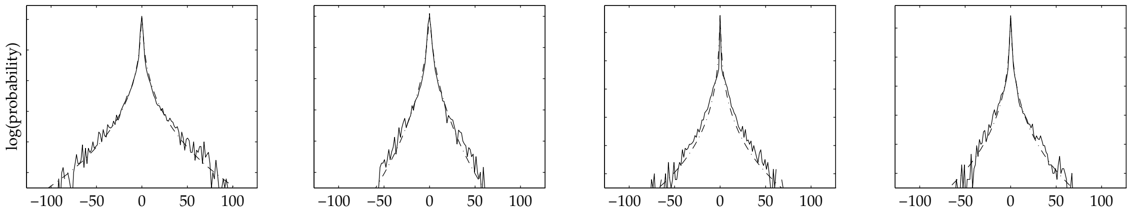
Example histograms[1] for wavelet subbands of several images are plotted in figure 1. Compared with a Gaussian, these densities are more sharply peaked at zero, with more extensive tails. The intuitive explanation is that images contain smooth areas interspersed with occasional sharp transitions (e.g., edges). The smooth regions produce small-amplitude coefficients, and the transitions produce sparse large-amplitude coefficients [11]. To quantify this, we give the sample kurtosis (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 that characterizes a Gaussian distribution.

Also shown in figure 1 are two-parameter density functions of the form [17, 28]:

$$\mathcal{P}(c) \propto e^{-|c/s|^p}. \qquad (1)$$

The density parameters $\{s, p\}$ are estimated by minimizing the relative entropy (the Kullback-Leibler divergence) between a discretized model distribution and the 256-bin coefficient histogram. The fits are generally quite good: The CTscan fit in figure 1 is the worst in our data set.

[1] By considering these as representative of the underlying coefficient densities, we are making an implicit assumption of spatial stationarity.



$\kappa = 14.5,\ \ \Delta H = 0.079$ $\kappa = 18.9,\ \ \Delta H = 0.067$ $\kappa = 22.9,\ \ \Delta H = 0.25$ $\kappa = 23.0,\ \ \Delta H = 0.076$

Figure 1. Examples of 256-bin coefficient histograms for vertical bands of four images ("Boats", "Lena", "CTscan", and "Toys"), plotted in the log domain. Also shown (dashed lines) are fitted model densities corresponding to equation (1). Below each histogram is the sample kurtosis (fourth moment divided by squared variance), and the relative entropy of the model.

Figure 2. 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.

2. Wavelet Joint Statistical Model

Despite the decorrelation properties of the wavelet decomposition, it is quite evident that wavelet coefficients are *not* statistically independent. Figure 2 shows the magnitudes (absolute values) of coefficients in a four-level separable wavelet decomposition. Large-magnitude coefficients tend to lie along ridges with orientation matching that of the subband. Large-magnitude coefficients also tend to occur at the same relative spatial locations in subbands at adjacent scales, and orientations.

In order to characterize these statistical relationships explicitly, consider two coefficients representing horizontal information at adjacent scales, but the same spatial location. Figure 3A shows the log-domain conditional histogram $\mathcal{H}\left(\log_2(C) | \log_2(P)\right)$, where P is the magnitude of the coarse-scale coefficient and C is the magnitude of the finer-scale coefficient. The right side of the distribution is unimodal and concentrated about a unit-slope line, indicating that C is roughly proportional to P. Furthermore, vertical cross sections (i.e., the histogram conditioned on a fixed value of P) have roughly the same shape for different values of P. The left side of the distribution is concentrated about a horizontal line, suggesting that C is independent of P in this region. The form of the histogram shown in figure 3A also holds for pairs of coefficients at adjacent spatial locations and orientations, and is surprisingly robust across a wide range of images [3].

Given the linear relationship between large-amplitude coefficients and the difficulty of characterizing the full density of a coefficient conditioned on its neighbors, we've examined a linear predictor for coefficient magnitude. Figure 3B shows a histogram of C conditioned on a linear combination of the magnitudes, Q_k, of eight adjacent coefficients in the same subband, two coefficients at other orientations, and a coefficient at a coarser scale. The linear combination, $\langle \vec{w}, \vec{Q} \rangle$, is chosen to be least-squares optimal. The histogram is similar to the single-band conditional histogram of figure 3A. But the linear region is extended, and the conditional variance is significantly reduced.

In order to determine which coefficients to include in the conditioning set $\{Q_k\}$, we calculated the mutual information between C and $\langle \vec{w}, \vec{Q} \rangle$ for a variety of choices of interband and intraband coefficients. We used a greedy algorithm to choose the causal conditioning neighbors yielding the largest reduction in entropy.

The form of these log-domain conditional histograms (in particular, the independence at small magnitudes, and the nearly constant shape that shifts linearly with the predictor at large magnitudes) suggests a model of multiplicative uncertainty with additive noise, which we write as:

$$C = M \cdot \langle \vec{w}, \vec{Q} \rangle + N, \tag{2}$$

where M and N are two mutually independent zero-mean random variables. The distribution of M is determined empirically. We construct a lookup table for the conditional cumulative distribution in the log domain by averaging the mean- and variance-normalized conditional histograms of three training images at two scales and all three orientations. We assume N is Gaussian distributed. The parameters of the model are thus the variances of M (in the log domain) and N, and the weights $\{w_k\}$. The weights are chosen to

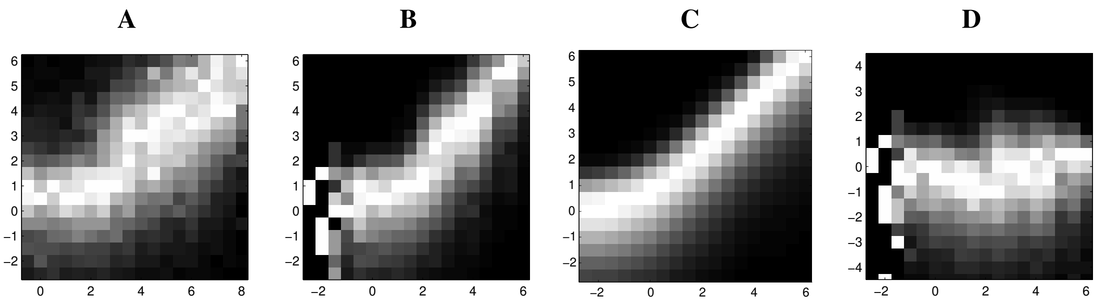


Figure 3. Conditional histograms of the log magnitude of a coefficient in a horizontal subband of the “Boats” image. Intensity corresponds to probability, except that each column has been independently rescaled to fill the full intensity range. **A**: Conditioned on the log magnitude of a coefficient in a coarser-scale horizontal subband. **B**: Conditioned on the log of a linear combination of coefficient magnitudes from adjacent spatial positions, orientations, and scales. **C**: Model of equation (2) fitted to the conditional histogram in B. **D**: Divisively normalized coefficient (see text).

be least-square optimal, and the variances are then chosen to minimize the relative entropy between the joint model density and the joint histogram. Figure 3C shows the model density that best fits the density of figure 3B.

Finally, consider the goal of independent components described in the introduction. The variance of coefficient C depends on its neighbors, $\vec{Q}$, but this dependency may be removed by calculating a *divisively normalized component*:

$$\tilde{C} = C/\sqrt{\langle \vec{w}, \vec{Q} \rangle^2 + \sigma^2}$$

Figure 3D shows the conditional density of this normalized component, which is seen to be roughly independent of $\langle \vec{w}, \vec{Q} \rangle$.

3. Compression

Wavelets have been particularly successful for image compression. Although many image coders do not incorporate an explicit probability model, a number of recent algorithms make use of joint statistical regularities between wavelet coefficients [19, 27, 22, 26, 15, 25, 6, 32, 16].

We have constructed two coders called EPWIC [4, 29, 3] based directly on the probability models described in sections 1 and 2. In both coders, subband coefficients are encoded one bitplane at a time using a non-adaptive arithmetic encoder that utilizes probabilities calculated from the corresponding 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.

Figure 4 shows a comparison of our coders to two well-known coders: the JPEG coder[2], and the Embedded Zerotree Wavelet (EZW) coder [27]. Also shown in figure 4 is the relative encoding size as a function of target PSNR. This gives a sense of how long one would wait during a progressive transmission for an image of a given quality. The

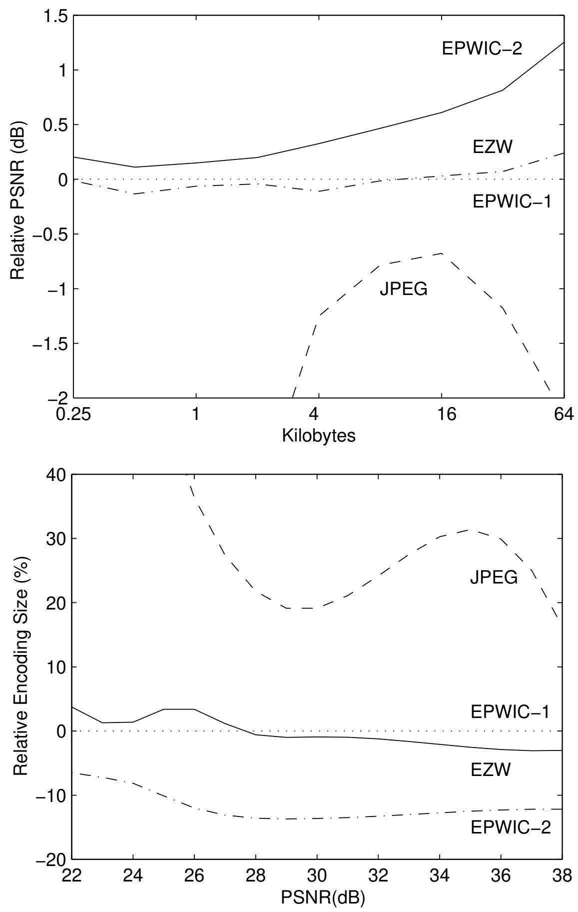


Figure 4. Relative rate-distortion tradeoff for four image coders (JPEG, EZW, EPWIC-1, and EPWIC-2). **Top**: PSNR values (in dB), relative to EPWIC-1, as a function of the number of encoded bytes. **Bottom**: Number of bytes necessary to achieve a given PSNR, relative to EPWIC-1. All curves are averages over a set of 13 images (512×512, 8-bit pixels).

[2] Independent JPEG Group’s CJPEG, version 5b.

coders perform quite well, despite considerable overhead, and a direct (non-heuristic) implementation. EPWIC-1 utilizes only marginal statistics, but is comparable to EZW, which captures coefficient structure across scale. EPWIC-2, which encodes the normalized components described in section 2, significantly outperforms the other techniques.

4. Restoration

The classical solution to the noise removal problem is the Wiener filter, which assumes an image model of independent Gaussian-distributed coefficients in the Fourier domain. On the other hand, the non-Gaussian marginal statistics of filtered subbands have been utilized implicitly in a noise removal procedure known as "coring" [e.g., 23, 1, 5, 18], a simple version of which is used in consumer videocassette players. In this approach, the image is split into subbands, and the coefficients are thresholded to suppress low-amplitude values while retaining high-amplitude values. Recent work has provided statistical justification for such algorithms [e.g., 10]. As a simple example, a least-squares Bayesian estimator, assuming non-Gaussian marginal statistics of the type shown in figure 1, is a nonlinear function in the form of a softened coring function [28].

We've been developing a semi-blind noise-removal algorithm based on the joint statistical model of section 2. We use an overcomplete tight frame representation with four oriented subbands at each scale [30]. We use a causal neighborhood of size 18 and bootstrap to estimate model parameters. For each subbands, from coarse to fine, we:

1. Adjust subband variances:

$$C' = \sqrt{\frac{\mathcal{E}(C^2)-\sigma^2}{\mathcal{E}(C^2)}} \cdot C$$

2. Estimate normalization weights based on C':

$$\vec{w} = \mathcal{E}\left(\vec{Q}\cdot\vec{Q}^T\right)^{-1}\cdot\mathcal{E}\left(\vec{Q}\cdot|C'|\right)$$

3. Compute a Wiener estimate, using the neighbor-predicted variance:

$$\hat{C} = \frac{\langle\vec{w},\vec{Q}\rangle^2}{\langle\vec{w},\vec{Q}\rangle^2+\sigma^2} \cdot C$$

Figure 5 shows comparisons of Wiener filtering, Bayes least-square coring [28], and the joint algorithm. Each algorithm is "semi-blind": the contaminating noise is additive white Gaussian with known variance. The joint algorithm produces better results, both visually and in SNR.

5. Synthesis

There is a large body of literature on texture synthesis, with much of it based on Markov Random Fields [e.g., 8, 12]. Recent techniques have been developed to synthesize images with the same wavelet coefficient marginal statistics as those of an example image. In particular, Heeger and Bergen [13] used an overcomplete basis, and iteratively alternated between matching the subband histograms, and matching the pixel histogram. Zhu et. al. [33] used Gibbs sampling to draw from the maximal-entropy distribution with the same marginals as the example image. This technique may be applied to the marginals of linear or non-linear operators. But in contrast to the Heeger and Bergen algorithm, it is extremely expensive computationally.

The use of joint statistics of rectified subband coefficients for texture analysis appears often in the human vision literature in the form of "second-order" texture models [e.g., 2, 7]. Recent nonlinear joint models have given impressive synthesis results. Popat and Picard [21] have developed a probability model for densities of local coefficient clusters (including those at different scales), and used it to synthesize texture examples. DeBonet and Viola [9] describe a fast heuristic synthesis technique which captures joint relationships across scale.

We've been working to develop a synthesis algorithm[3] based on the joint statistical observations of section 2. Within each subband of a steerable pyramid [30], we match the correlation of the magnitudes of a set of neighboring (in spatial position, orientation, and scale) coefficients, $\vec{Q}$, using a linear transformation closest to the identity. In particular, we solve for matrix A that

$$\begin{aligned} \text{minimizes:} \quad & \mathcal{E}\left(||\vec{Q}-A\vec{Q}||^2\right) \\ \text{subject to:} \quad & \mathcal{E}\left(A\vec{Q}\vec{Q}^TA^T\right) = \mathcal{E}\left(\vec{Q_0}\vec{Q_0}^T\right), \end{aligned}$$

We alternate between applying A, and projecting into the image subspace. The density from which we are sampling is thus not explicit, but the algorithm does converge quickly (minutes). Figure 6 shows comparisons to the Heeger-Bergen algorithm for three example textures. For each, the joint algorithm is seen to better capture the structure of the underlying texture. The second example (a piece of fabric) is particularly impressive, as our algorithm captures both the fine-scale diagonals of the fabric, and the organization of these diagonals into vertical bands.

6. Discussion

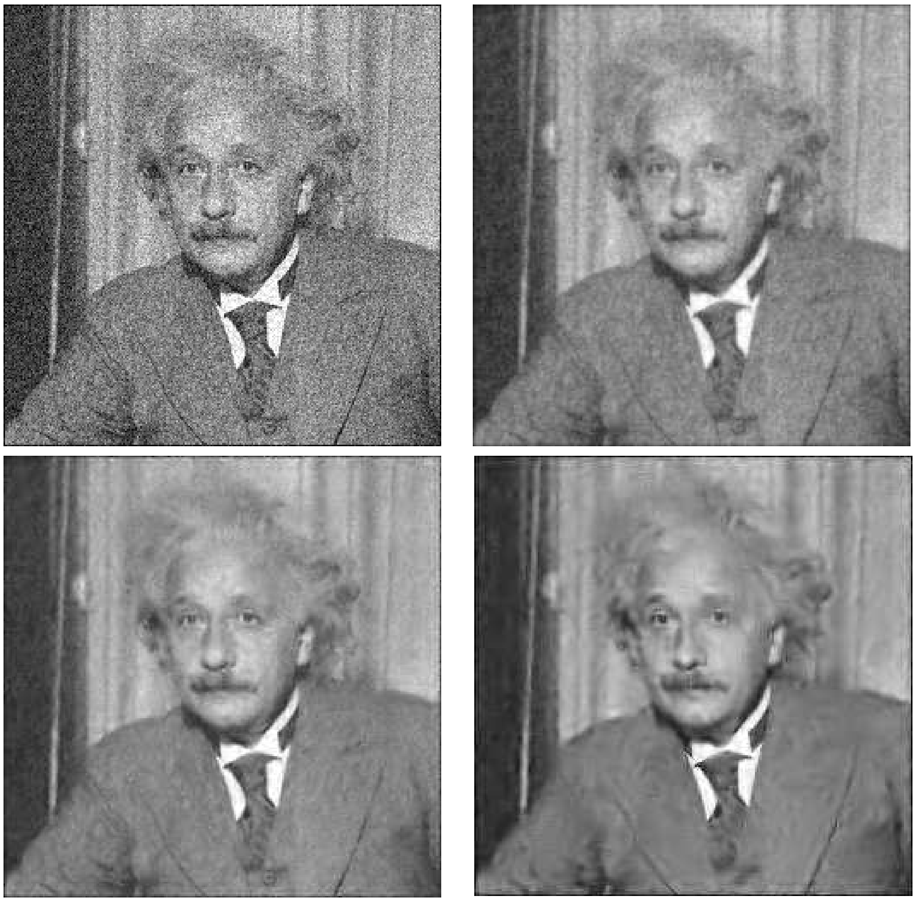
We've described a parameterized model for the joint statistics of wavelet coefficient magnitudes, and demonstrated its use in applications of compression, restoration, and synthesis. The results are quite strong, considering the simplicity of the model.

Many aspects of the model could be improved. The most obvious of these is to describe the *signs* of the coefficients, which exhibit significant statistical regularity. We estimate

[3] Joint work with Javier Portilla, Instituto de Optica.

Noisy
4.78 dB

Wiener
9.79 dB



Coring
11.67 dB

Joint Coring
12.71 dB

Figure 5: Semi-blind noise removal example (see text).

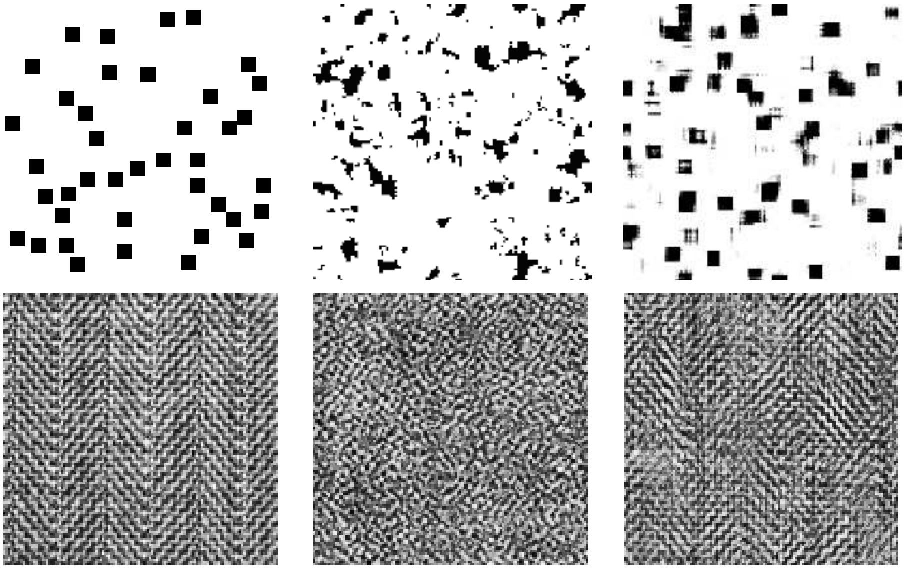


Figure 6. Texture synthesis examples. **Left**: Original texture example. **Center**: First-order model synthesis [13]. **Right**: Joint model synthesis.

that the entropy of the sign bits, given a simple prediction from neighboring signs, is about 0.4 bits/coefficient and could thus give substantial coding improvements. These relationships could also improve the restoration and synthesis results. In addition, the model does not accurately account for statistics of overcomplete bases (as used in both the restoration and synthesis examples), since the coefficients of such transforms are not decorrelated.

There are also many other interesting applications. We are working to develop more controlled synthesis and enhancement techniques that can take into account spatial or scale constraints [e.g., 14]. A simple example is "super-resolution", in which one synthesizes fine-scale detail. In addition, the decomposition into independent components (described above) using divisive normalization may provide the first theoretical justification for current models of processing in visual cortex [e.g., 31].

References

[1] B. E. Bayer and P. G. Powell. A method for the digital enhancement of unsharp, grainy photographic images. *Adv in Computer Vision and Im Proc*, 2:31–88, 1986.

[2] J. R. Bergen and E. H. Adelson. Early vision and texture perception. *Nature*, 333:363–364, 1988.

[3] R. W. Buccigrossi and E. P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. Tech Report 414, GRASP Lab, U. Pennsylvania, May 1997.

[4] R. W. Buccigrossi and E. P. Simoncelli. Progressive wavelet image coding based on a conditional probability model. In *ICASSP*, Munich, Apr 1997.

[5] C. Carlson, E. Adelson, and C. Anderson. Improved system for coring an image representing signal, 1985. US Patent 4,523,230.

[6] C. Chrysafis and A. Ortega. Efficient context-based entropy coding for lossy wavelet image coding. In *Data Compression Conf*, Snowbird, Utah, Mar 1997.

[7] C. Chubb and G. Sperling. Drift-balanced random stimuli: A general basis for studying non-fourier motion perception. *JOSA A*, 5:1986–2007, 1988.

[8] G. Cross and A. Jain. Markov random field texture models. *IEEE Trans PAMI*, 5:25–39, 1983.

[9] J. De Bonet and P. Viola. A non-parametric multi-scale statistical model for natural images. In *Adv in Neural Info Processing*, Dec 1997. To Appear.

[10] D. Donoho and I. Johnstone. Adapting to unknown smootness via wavelet shrinkage. *J Am Stat Assoc*, 90(432), Dec 1995.

[11] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J Opt Soc Am A*, 4(12):2379–2394, Dec 1987.

[12] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans PAMI*, 6:721–741, 1984.

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

[14] A. N. Hirani and T. Totsuka. Combining frequency and spatial domain information for fast interactive image noise removal. In *ACM SIGGRAPH*, pp 269–276, 1996.

[15] F. Kossentini, W. Chung, and M. Smith. A jointly optimized subband coder. *IEEE Trans Im Proc*, 5(9):1311–1323, Sep 1996.

[16] S. M. LoPresto, K. Ramchandran, and M. T. Orchard. Wavelet image coding based on a new generalized gaussian mixture model. In *Data Compression Conf*, Snowbird, Utah, Mar 1997.

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

[18] J. M. Ogden and E. H. Adelson. Computer simulations of oriented multiple spatial frequency band coring. Tech Report PRRL-85-TR-012, RCA David Sarnoff Research Center, Apr 1985.

[19] 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.

[20] A. P. Pentland. Fractal based description of natural scenes. *IEEE Trans PAMI*, 6(6):661–674, 1984.

[21] K. Popat and R. W. Picard. Cluster-based probability model and its application to image and texture processing. *IEEE Trans Im Proc*, 6(2):268–284, Feb 1997.

[22] R. Rinaldo and G. Calvagno. Image coding by block prediction of multiresolution subimages. *IEEE Trans Im Proc*, Jul 1995.

[23] J. P. Rossi. *JSMPTE*, 87:134–140, 1978.

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

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

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

[27] J. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans Sig Proc*, 41(12):3445–3462, Dec 1993.

[28] E. P. Simoncelli and E. H. Adelson. Noise removal via bayesian wavelet coring. In *Int'l Conf on Im Proc*, pp 379–383, Lausanne, Sep 1996.

[29] E. P. Simoncelli and R. W. Buccigrossi. Embedded wavelet image compression based on a joint probability model. In *Int'l Conf on Im Proc*, Santa Barbara, Oct 1997.

[30] E. P. Simoncelli and W. T. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Int'l Conf on Im Proc*, Washington, Oct 1995.

[31] E. P. Simoncelli and D. J. Heeger. A model of neuronal responses in visual area MT. *Vision Research*, 1997. In press.

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

[33] S. Zhu, Y. Wu, and D. Mumford. Filters, random fields and maximum entropy (FRAME) – towards the unified theory for texture modeling. In *IEEE Conf Comp Vision and Patt Rec*, Jun 1996.