Published in: Proc. SPIE 44th Annual Meeting, vol. 3813, pp. 188-195,
Denver, Colorado. July 1999.

Modeling the Joint Statistics of Images in the Wavelet Domain

Eero P. Simoncelli

Center for Neural Science, and
Courant Institute of Mathematical Sciences
New York University
4 Washington Place, Room 809
New York, NY 10012

ABSTRACT

I describe a statistical model for natural photographic images, when decomposed in a multi-scale wavelet basis. In particular, I examine both the marginal and pairwise joint histograms of wavelet coefficients at adjacent spatial locations, orientations, and spatial scales. Although the histograms are highly non-Gaussian, they are nevertheless well described using fairly simple parameterized density models.

Keywords: image model, wavelet, statistical inference, non-Gaussian statistics

The set of visual images is enormous, and yet only a small fraction of these are likely to be encountered in a natural setting. Thus, a statistical prior model, even one that only partially captures these variations in likelihood, can substantially benefit image processing and artificial vision systems. The problem of inferring a probability density for images is difficult because of their high dimensionality. To make progress, it is thus essential to reduce the dimensionality of the space on which one defines the probability model.

Two basic assumptions are commonly used for this purpose. The first is a Markov assumption: the probability density of a pixel*, when conditioned on a set of pixels in a small spatial neighborhood, is independent of the pixels beyond the neighborhood. The second is an assumption of translation-invariance (i.e., strict-sense stationarity): the distribution of pixels in a neighborhood does not depend on the absolute location of that neighborhood within the image. Markov random field models have been widely used for texture modeling [e.g., 3–5]. They are, however, inadequate for image modeling because images can contain structures that are arbitrarily large and that thus extend beyond any small neighborhood of pixels.

The power of these statistical models may be substantially augmented if the image is first transformed to a new space, in which the density has a more local structure. Choosing a basis that is appropriate for the statistics of an input signal is a classical problem. The traditional solution is principal components analysis (PCA), in which a linear decomposition is chosen to decorrelate the input signal. If one assumes that the model should be strict-sense stationary, then the Fourier transform will produce such diagonalization. The most widely used model of image statistics is a Fourier model, in which the power spectral density falls inversely as a power of spatial frequency [e.g., 6–9].

It is well known that the PCA transform is not unique. First, if the covariance matrix has non-distinct eigenvalues, one can arbitrarily rotate within the associated subspace without affecting the covariance structure of the signal. Second, after transforming the signal to the PCA coordinate system and scaling the axes to unit variance (i.e., "whitening" the input), one can rotate the entire space without affecting the covariance structure. Recent work on "independent components analysis" (ICA) uses higher-order statistical measures in order to uniquely constrain the linear basis [e.g., 10,11]. In particular, several authors have constructed optimal bases for images by optimizing such measurements [12,13]. The resulting basis functions are spatially oriented and have spatial frequency bandwidths of roughly one octave, similar to the most common multi-scale decompositions. These basis functions have also been compared with the "receptive field" properties of neurons in mammalian visual cortex.

E-mail: eero.simoncelli@nyu.edu. Some of the material in this paper has been previously published in [1,2].

*I'll assume the image has been discretely sampled.

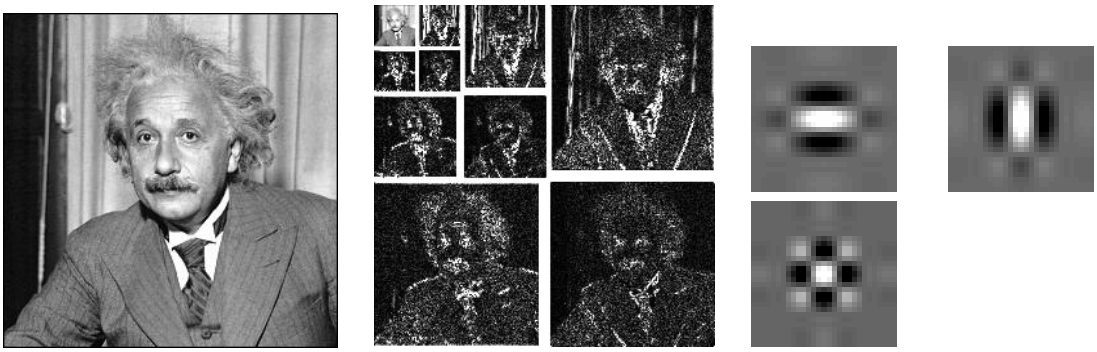

Figure 1. Coefficient magnitudes of a wavelet decomposition. Left: original "Einstein" image. Middle: Absolute values of subband coefficients at three scales, and three orientations of a separable wavelet decomposition of this image. Right: example basis functions (three different "orientations", at a single coarse scale).

Our experience is that these observations of highly kurtotic marginals do not depend critically on the choice of basis, beyond the qualitative properties of orientation specificity and roughly octave bandwidth. For our purposes in this paper, we utilize a fixed multi-scale orthonormal wavelet basis, in which functions are related by translation, dilation, and frequency modulation.[†] The image is decomposed into a set of *subbands*, each consisting of those coefficients associated with a set of basis functions related by translation (but all having the same orientation and scale). An example of the decomposition is shown in figure 1, along with example basis functions at a single scale.

1. MARGINAL MODEL

Once we've transformed the image to wavelet domain, how do we characterize the statistical properties of the coefficients? Following the discussion above, the simplest model assumes the coefficients within a subband are independent and identically distributed. The model is thus completely determined by the marginal statistics of the coefficients.

Empirically, the marginal statistics of wavelet coefficients of natural images are highly non-Gaussian [e.g., 7,15]. In particular, the histograms are found to have much heavier tails and are more sharply peaked than one would expect from a Gaussian density. As an example, histograms[‡] of separable wavelet decompositions of several images are plotted in figure 2.

These marginal densities are well-modeled by a generalized Laplacian (or "stretched exponential") distribution [e.g., 15,16]:

$$\mathcal{P}_c(c; s, p) = \frac{e^{-|c/s|^p}}{Z(s,p)}, \tag{1}$$

where the normalization constant is $Z(s,p) = 2\frac{s}{p}\Gamma(\frac{1}{p})$. Each graph in figure 2 includes a dashed curve corresponding to the best fitting instance of this density function, with the parameters $\{s, p\}$ estimated by maximizing the likelihood of the data under the model. We have observed that values of the exponent p typically lie in the range $[0.5, 0.8]$. The factor s varies monotonically with the size of the basis functions, producing higher variance for coarser-scale components. The density model fits the histograms remarkably well, as indicated by the relative entropy measures given below each plot.

Figure 3 illustrates the entropy gain associated with this model, as compared to a Gaussian density model, and an empirical density model (i.e., assuming a known histogram). The Gaussian model variance was set to

[†]More precisely, the results shown in this paper are based on a 9-tap quadrature mirror filter (QMF) decomposition, using filters designed in [14]. This is a linear-phase (symmetric) approximation to an orthonormal wavelet decomposition. The drawback is that the system does not give perfect reconstruction: The filters are designed to optimize a residual function. This is not a serious problem for most applications, since reconstruction signal-to-noise ratios (SNRs) are typically about 55dB.

[‡]Histograms are obtained by gathering the coefficients of a single subband. Thus, we assume ergodicity and strict-sense stationarity in order to consider them as representative of the underlying density.

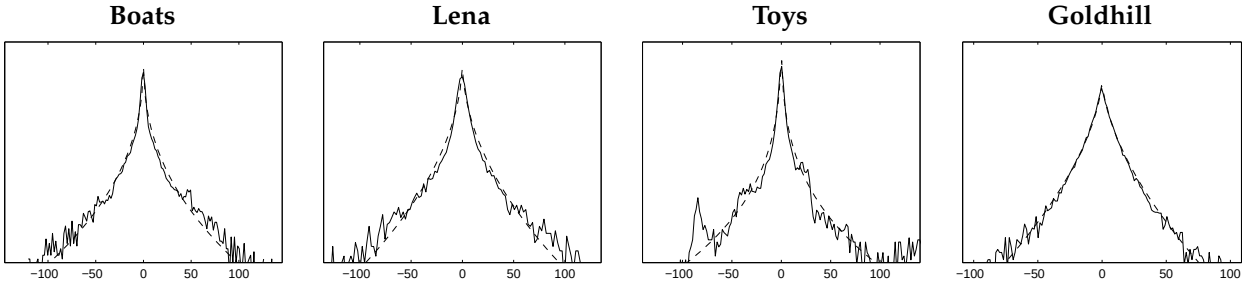

Figure 2. Examples of 256-bin coefficient histograms for a single vertical wavelet subband of several images, plotted in the log domain. All images are size $512x512$. Also shown (dashed lines) are fitted model densities corresponding to equation (1). Below each histogram is the maximum-likelihood value of p used for the fitted model density, and the relative entropy (Kullback-Leibler divergence) of the model and histogram, as a fraction of the total entropy of the histogram.

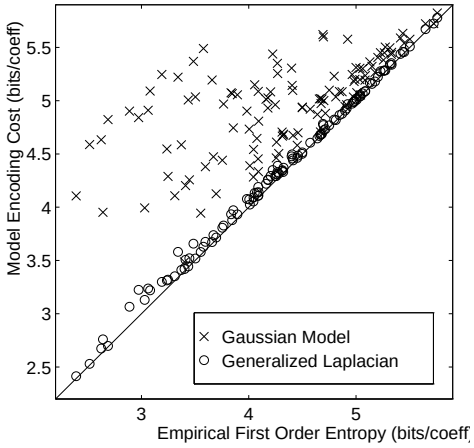

Figure 3. Comparison of encoding costs. Plotted are costs of encoding data from subbands assuming the generalized Laplacian density of equation (1) (O's), and the encoding cost assuming a Gaussian density (X's), versus the empirical encoding cost using a 256-bin histogram. Points are plotted for 9 bands (3 scales, 3 orientations) of 13 images.

match the sample variance. Note that the generalized Laplacian model comes within 0.25 bits/coefficient of the empirical entropy, as compared with the Gaussian density model which often has a relative entropy greater than 1.0 bit/coefficient.

Thus, our first model of image statistics is that each wavelet subband consists of independent identically distributed random variables drawn from a density of the form given in equation (1). The model is parameterized by the set $\{s_k, p_k\}$, where k indexes the subband. Since the wavelet transform is orthonormal, we can easily draw statistical samples from this model. Figure 4 shows the result of drawing the coefficients of a wavelet representation independently from generalized Laplacian densities. The density parameters for each subband were chosen as those that best fit the "Einstein" image. Although it has more structure than an image of white noise, the result does not look very much like a photographic image!

2. JOINT MODEL

The primary reason for the poor appearance of the image in figure 4 is that the coefficients of the wavelet transform are not independent. Intuitively, independence would seem unlikely, since images are not formed from linear superpositions of independent patterns. In fact, the typical combination rule for image formation is *occlusion*.

Empirically, the coefficients of orthonormal wavelet decompositions of visual images are found to be fairly well decorrelated (i.e., their covariance is zero). They are *not*, however, independent [2,1]. One can *see* this in

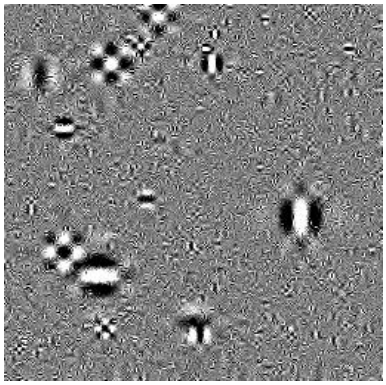

Figure 4. An image drawn from the subband marginal model, with subband density parameters chosen to fit the image of figure 1.

figure 1: the middle panel shows the magnitudes (absolute values) of coefficients in a four-level separable wavelet decomposition. Note that large-magnitude coefficients tend to occur near each other within subbands, and also occur at the same relative spatial locations in subbands at adjacent scales, and orientations [e.g., 2,1].

To make this more explicit, the top row of figure 5 shows joint histograms of several different pairs of coefficients. As with the marginals, we assume ergodicity and strict-sense stationarity in order to consider the joint histogram of this pair of coefficients, gathered over the spatial extent of the image, as representative of the underlying density. The adjacent coefficients produce contours that are nearly circular, whereas the others are clearly extended along the axes. As a side note, the concentration of the probability mass along the axes suggests that the wavelet decomposition is a reasonable choice of transform (at least for this particular image). Zetzsche [17] has examined at empirical joint densities of quadrature (Hilbert transform) pairs of basis functions and found that the contours are roughly circular.

The joint histograms shown in the first row of figure 5 do not make explicit the issue of whether the coefficients are independent. The bottom row of shows *conditional* histograms of the same data. Let c correspond to the density coefficient (vertical axis), and p the conditioning coefficient (horizontal axis). The histograms illustrate several important aspects of the relationship between the two coefficients. First, they are fairly well second-order decorrelated, since the expected value of c is approximately zero for all values of p. Second, the variance of the conditional histogram of c clearly depends the value of p, and the strength of this dependency depends on the particular pair of coefficients being considered. Thus, although c and p are uncorrelated, they still exhibit statistical dependence!

We have found that re-plotting the histograms as a function of the log coefficients reveals a simple structure within the data [2,1]. Figure 6A shows the conditional histogram $\mathcal{H}\left(\log_2(c^2)|\log_2(p^2)\right)$ of a "child" coefficient c conditioned on a coarser-scale "parent" coefficient p. The right side of the distribution is unimodal and concentrated along a unit-slope line, suggesting that the conditional expectation, $\mathbf{E}(\mathsf{c}^2|\mathsf{p}^2)$, is approximately proportional to p^2. Furthermore, vertical cross sections have approximately the same shape for different values of p^2. Finally, the left side of the distribution is concentrated about a horizontal line, suggesting that c^2 is independent of p^2 in this region.

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.

A simple simulation confirms this intuition. We computed the joint density of horizontal wavelet coefficient magnitudes at two adjacent scales for a 64×64 pixel image of a disk. The disk radius was drawn uniformly from the interval $[8, 24]$ pixels, and its position was randomized within ± 8 pixels of the image center. The disk amplitude was drawn from a generalized Laplacian density, with parameters $s = 1, p = 0.5$. We also added a small amount (SNR = 30 dB) of uniformly distributed white noise to the image. We collected statistics over 400 such images. The resulting distribution, plotted in figure 6B, is qualitatively similar to that of figure 6A.

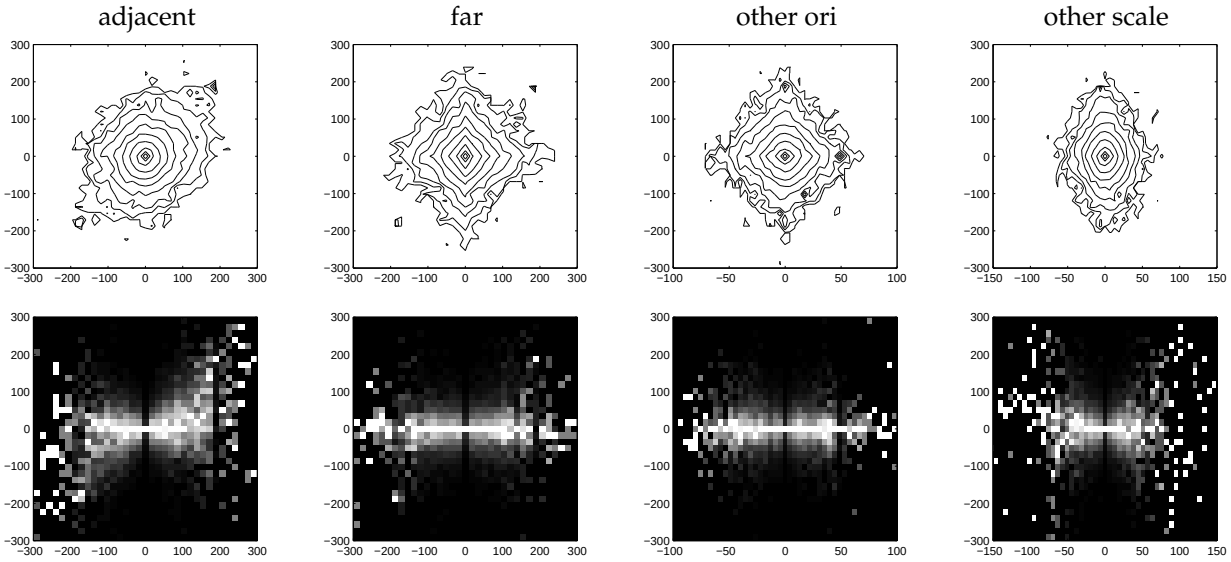

Figure 5. Empirical joint distributions of wavelet coefficients associated with different pairs of basis functions. The top row shows joint distributions as contour plots, with lines drawn at equal intervals of log probability. The two leftmost examples correspond to pairs of basis functions at the same scale and orientation, but separated by different spatial offsets. The third corresponds to a pair at orthogonal orientations (but the same scale and nearly the same position), and the rightmost corresponds to a pair at adjacent scales (but the same orientation, and nearly the same position). The bottom row shows corresponding conditional distributions: brighness corresponds to probability, except that each column has been independently rescaled to fill the full range of intensities. pStatistics were gathered (spatially) from the "Mountain" image.

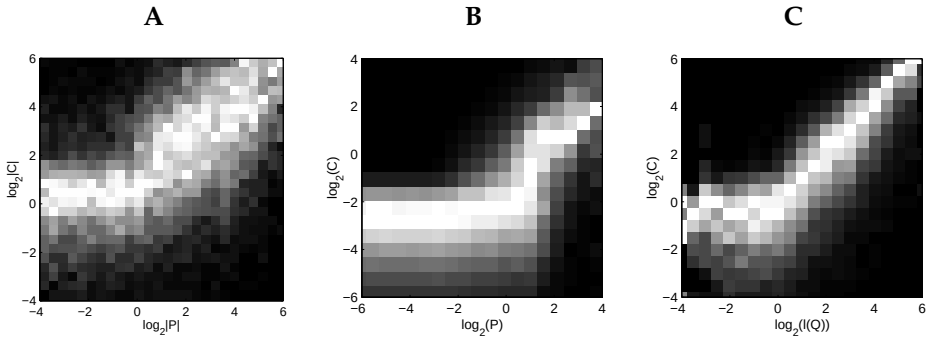

Figure 6. Conditional histograms for the log magnitude of a fine scale horizontal coefficient. Brightness corresponds to probability, except that each column has been independently rescaled to fill the full range of display intensities. **A:** Conditioned on the Parent (same location and orientation, coarser scale) coefficient. Data are for the "Boats" image. **B:** Same, except data are drawn from synthetic images containing non-overlapping disks of randomized spatial position and size and additive noise (see text). **C** Conditioned on a linear combination of neighboring coefficient magnitudes. Data are for the same subband of the Boats image as **A**.

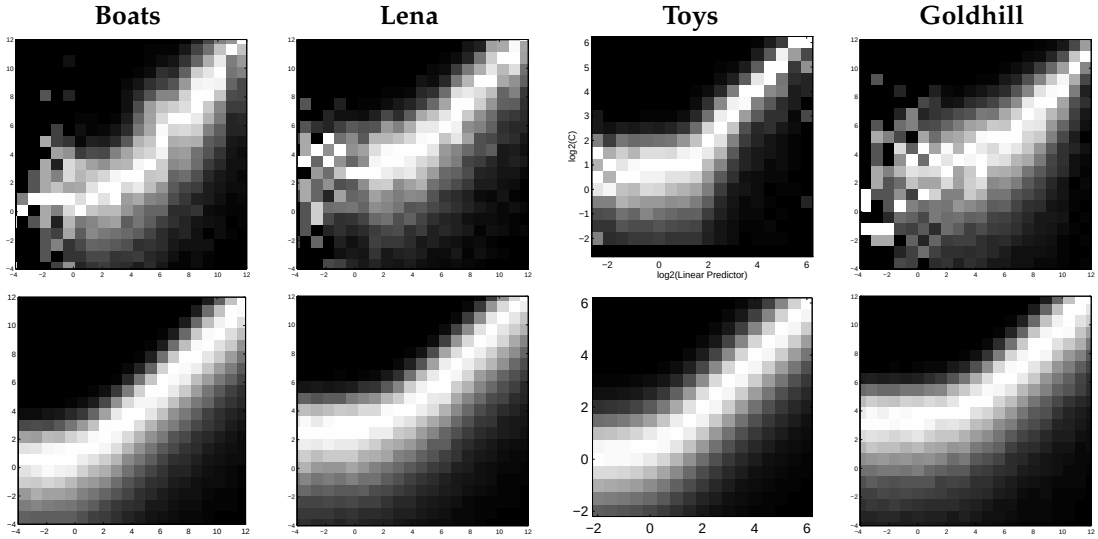

Figure 7. Top: Examples of log-domain conditional histograms for the second-level horizontal subband of different images, conditioned on an optimal linear combination of coefficient magnitudes from adjacent spatial positions, orientations, and scales. **Bottom**: Model of equation (2) fitted to the conditional histograms. Intensity corresponds to probability, except that each column has been independently rescaled to fill the full range of intensities.

We have modeled the conditional relationship between the two coefficients as:

$$\mathcal{P}(c|p) \;\propto\; f(c/\sqrt{p^2+\sigma^2}).$$

That is, the distribution of c is described by a density function f, with variance proportional to $p^2+\sigma^2$.

The form of the histograms shown in figure 6 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, adjacent orientations. As one considers coefficients that are more distant (either in spatial position or in scale), the dependency becomes weaker. This suggests that a Markov assumption might be appropriate.

Given the difficulty of characterizing the full density of a coefficient conditioned on a large set of neighbors, we decided to simply extend the variance-scaling relationship described above using a weighted linear predictor for the coefficient variance. This is a somewhat *ad hoc* choice, but it has proven effective in a number of applications of this model. In particular, we assume the variance of a coefficient, c scales as a linear combination of the squared coefficients in a local neighborhood:

$$\mathcal{P}(c \mid \{p_k\}) \;\propto\; f\left(c/\sqrt{\sum_k w_k p_k^2 + \sigma^2}\right). \tag{2}$$

Unlike traditional Gauss-Markov random fields, in which the conditional mean of a random variable depends linearly on the neighboring variables, here the *variance* of c depends linearly on the *squared* neighbors. If one assumes a particular form for the density f (e.g., Gaussian), the parameters $\{w_k, \sigma\}$ may be determined (numerically) via maximum likelihood estimation.

The model fits the histograms quite well, as can be seen from the examples shown in figure 7. The benefit of the model, in terms of entropy reduction, is summarized in figure 8.[§] The entropy savings are less than those illustrated in figure 3, but still substantial.

Sampling from this model is unfortunately not as straightforward as it is for the marginal model. We have been investigating the use of Monte Carlo and iterative techniques for this purpose [18].

[§]These entropy calculations were performed for a related model, in which the conditional standard deviation of the coefficient was modeled as a weighted sum of magnitudes of neighboring coefficients.

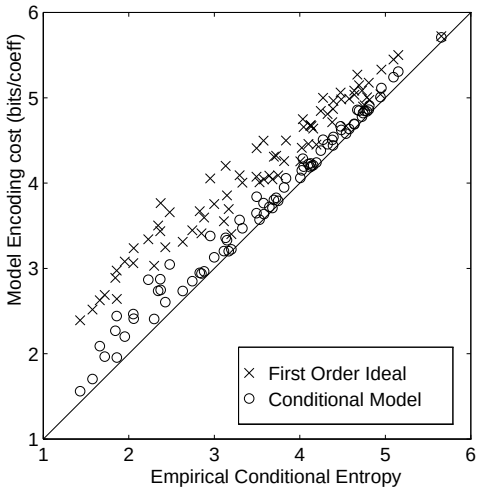

Figure 8. Comparison of encoding cost using the conditional probability model of equation (2) with the encoding cost using the first-order histogram, as a function of the encoding cost using a 256×256-bin joint histogram. Points are plotted for 6 bands (2 scales, 3 orientations) of 13 images.

3. DISCUSSION

I've described non-Gaussian marginal and joint models for visual images in the wavelet domain. The models are substantially more powerful than traditional Gaussian models. Our previous work has demonstrated that these types of probability model are quite powerful when used in applications of compression [2], denoising [16,19], and texture synthesis [20].

Although the empirical observations used to motivate the two models are quite striking, there are difficult issues to be resolved, especially in the precise nature of the joint model. Specifically, the observations are all of *pairs* of coefficients, but we have assumed a particular form for the conditional density based on a neighborhood of many coefficients. We have recently begun to investigate the use of scale mixtures for modeling wavelet coefficients [21]. This appears quite promising, as it gives a more rigorous justification to conditional variance behavior, and simplifies the problem of sampling.

Another difficult issue is that of translation-invariance. Orthonormal wavelet transform are not translation invariant: more precisely, the basis functions associated with a single subband do not span a translation-invariant subspace [22]. Thus, our model is not translation-invariant. Using an overcomplete multi-scale transforation can restore translation-invariance, but results in linear dependencies between the coefficients [e.g. 19]. For some applications, results may be averaged over all translates of the input image, in a process known as "cycle spinning" [23]. Alternatively, one can adaptively "sparsify" the representation [e.g., 24,25], but this is, in general, a global non-convex optimization problem.

ACKNOWLEDGMENTS

Funding for this research was provided by NSF CAREER grant MIP-9796040.

REFERENCES

1. E. P. Simoncelli, "Statistical models for images: Compression, restoration and synthesis," in *31st Asilomar Conf on Signals, Systems and Computers*, pp. 673–678, IEEE Computer Society, (Pacific Grove, CA), November 1997.
2. R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," Tech. Rep. 414, GRASP Laboratory, University of Pennsylvania, May 1997. Accepted (3/99) for publication in IEEE Trans Image Processing.
3. M. Hassner and J. Sklansky, "The use of Markov random fields as models of texture," *Comp. Graphics Image Proc.* **12**, pp. 357–370, 1980.

4. R. L. Kashyap, R. Chellappa, and A. Khotanzad, "Texture classification using features derived from random field models," *Patt Rec Letters* **1**, pp. 43–50, Oct 1982.
5. G. Cross and A. Jain, "Markov random field texture models," *IEEE Trans PAMI* **5**, pp. 25–39, 1983.
6. A. Pentland, "Fractal based description of natural scenes," *IEEE Trans. PAMI* **6**(6), pp. 661–674, 1984.
7. D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Am. A* **4**(12), pp. 2379–2394, 1987.
8. D. L. Ruderman and W. Bialek, "Statistics of natural images: Scaling in the woods," *Phys. Rev. Letters* **73**(6), 1994.
9. A. van der Schaaf and J. H. van Hateren, "Modelling the power spectra of natural images: Statistics and information," *Vision Research* **28**(17), pp. 2759–2770, 1996.
10. J. F. Cardoso, "Source separation using higer order moments," in *ICASSP*, pp. 2109–2112, 1989.
11. P. Common, "Independent component analysis, a new concept?," *Signal Process.* **36**, pp. 387–314, 1994.
12. B. A. Olshausen and D. J. Field, "Natural image statistics and efficient coding," *Network: Computation in Neural Systems* **7**, pp. 333–339, 1996.
13. A. J. Bell and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision Research* **37**(23), pp. 3327–3338, 1997.
14. E. P. Simoncelli and E. H. Adelson, "Subband transforms," in *Subband Image Coding*, J. W. Woods, ed., ch. 4, pp. 143–192, Kluwer Academic Publishers, Norwell, MA, 1990.
15. S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Pat. Anal. Mach. Intell.* **11**, pp. 674–693, July 1989.
16. E. P. Simoncelli and E. H. Adelson, "Noise removal via Bayesian wavelet coring," in *Third Int'l Conf on Image Proc*, vol. I, pp. 379–382, IEEE Sig Proc Society, (Lausanne), September 1996.
17. C. Zetzsche, B. Wegmann, and E. Barth, "Nonlinear aspects of primary vision: Entropy reduction beyond decorrelation," in *Int'l Symposium, Society for Information Display*, vol. XXIV, pp. 933–936, 1993.
18. E. Simoncelli and J. Portilla, "Texture characterization via joint statistics of wavelet coefficient magnitudes," in *Fifth IEEE Int'l Conf on Image Proc*, vol. I, IEEE Computer Society, (Chicago), October 4-7 1998.
19. E. P. Simoncelli, "Bayesian denoising of visual images in the wavelet domain," in *Bayesian Inference in Wavelet Based Models*, P. Müller and B. Vidakovic, eds., Springer-Verlag, New York, June 1999. Lecture Notes in Statistics 141.
20. J. Portilla and E. P. Simoncelli, "Texture representation and synthesis using correlation of complex wavelet coefficient magnitudes," Tech. Rep. 54, Consejo Superior de Investigaciones Cientificas (CSIC), Madrid, March 1999. Available at ftp://ftp.cns.nyu.edu/pub/eero/portillaTR54-ltr.ps.gz. Submitted to Int'l Journal on Computer Vision.
21. M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," in *NIPS*99*, December 1999. Submitted, 21 May 1999.
22. E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans Information Theory* **38**, pp. 587–607, March 1992. Special Issue on Wavelets.
23. R. R. Coifman and D. L. Donoho, "Translation-invariant de-noising," Tech. Rep. 475, Statistics Department, Stanford University, May 1995.
24. D. L. Donoho and I. M. Johnstone, "Ideal denoising in an orthogonal basis chosen from a library of bases," *C.R. Acad. Sci.* **319**, pp. 1317–1322, 1994.
25. M. Lewicki and B. Olshausen, "Inferring sparse, overcomplete image codes using an efficient coding framework," in *Proc NIPS*97*, pp. 815–821, MIT Press, May 1998.