4.7

Statistical Modeling of Photographic Images

Eero P. Simoncelli
New York University

1 Introduction

The set of all possible visual images is huge, but not all of these are equally likely to be encountered by an imaging device such as the eye. Knowledge of this nonuniform probability on the image space is known to be exploited by biologic visual systems and can be used to advantage in most applications in image processing and machine vision. For example, loosely speaking, when one observes a visual image that has been corrupted by some sort of noise, the process of estimating the original source image may be viewed as one of looking for the highest-probability image that is "close to" the noisy observation. The problem of compression essentially boils down to using a larger proportion of the available bits to encode those regions of the image space that are more likely. And, problems such as resolution enhancement or image synthesis involve selecting (sampling) a high-probability image from the distribution, perhaps subject to some set of constraints. Precise descriptions of such applications can be found in many chapters throughout this book.

To develop a probability model for visual images, we first must decide which images to model. In a practical sense, this means we must (a) decide on imaging conditions, such as the field of view, resolution, sensor or postprocessing nonlinearities, and so forth, (b) decide what kind of scenes, under what kind of lighting, are to be captured in the images. It may seem odd, if one has not encountered such models, to imagine that all images are drawn from a single universal probability urn. In particular, the features and properties in any given image are often specialized. For example, outdoor nature scenes contain structures that are quite different from city streets, which in turn are nothing like human faces. There are two means by which this dilemma is resolved. First, the statistical properties that we will examine are basic enough that they are relevant for essentially *all* visual scenes. Second, we will use parametric models, in which a set of hyperparameters (possibly random variables themselves) govern the detailed behavior of the model, and thus allow a certain degree of adaptability of the model to different types of source material.

How does one build and test a probability model for images? Many approaches have been developed, but in this chapter, we'll describe an empirically driven methodology based on the study of discretized (pixelated) images. Currently available digital cameras generate such images, typically containing millions of pixels. Naively, one could imagine examining a large set of such images to try to determine how they are distributed. But a moment's thought leads one to realize the hopelessness of the endeavor. The amount of data needed to estimate a probability distribution from samples grows as K^D, where D is the dimensionality of the space (in this case, the number of pixels). This is known as the "curse of dimensionality."

Thus, to make progress on image modeling, it is *essential* to reduce the dimensionality of the space. Two types of simplifying assumption can help in this regard. The first, known as a Markov assumption, is that the probability density of a pixel,

when conditioned on a set of pixels in a small spatial neighborhood, is independent of the pixels outside of the neighborhood. A second type of simplification comes from imposing symmetries or invariances on the probability structure. The most common of these is that of translation-invariance (sometimes called homogeneity, or strict-sense stationarity): The distribution of pixels in a neighborhood does not depend on the absolute location of that neighborhood within the image. This seems intuitively sensible, given that a lateral or vertical translation of the camera leads approximately to a translation of the image intensities across the pixel array. Note that translation-invariance is not well defined at the boundaries, and as is often the case in image processing, these locations must usually be handled specially.

Another common assumption is scale-invariance: Resizing the image does not alter the probability structure. This may also be loosely justified by noting that adjusting the focal length (zoom) of a camera lens approximates (apart from perspective distortions) image resizing. As with translation-invariance, scale-invariance will clearly fail to hold at certain "boundaries." Specifically, scale-invariance must fail for discretized images at fine scales approaching the size of the pixels. And, similarly, it must also fail at coarse scales approaching the size of the entire image.

With these sort of simplifying structural assumptions in place, we can return to the problem of developing a probability model. In recent years, researchers from image processing, computer vision, physics, applied math, and statistics have proposed a wide variety of different types of model. In this chapter, I will review some basic statistical properties of photographic images, as observed empirically, and describe several models that have been developed to incorporate these properties. I will give some indication of how these models have been validated by examining how well they fit the data, but the true test usually comes when one uses the model to solve an image processing problem (such as compression or denoising). Although this is somewhat beyond the scope of this chapter, I will show some simple denoising examples to give an indication of how much performance gain one can obtain by using a better statistical model. To keep the discussion focused, I will limit the discussion to discretized *gray-scale* photographic images. Many of the principles are easily extended to color photographs [8, 43], temporal image sequences (movies) [15], and more specialized image classes such as portraits, landscapes, or textures. In addition, the general approach may also be applied to nonvisual imaging devices, such as medical images, infrared images, radar and other types of range image, or astronomic images.

2 The Gaussian Model

The classic model of image statistics was developed by television engineers in the 1950s (see [41] for a review), who were interested in optimal signal representation and transmission. The most basic motivation for these models comes from the observation that pixels at nearby locations tend to have similar intensity values. This is easily confirmed by measurements like those shown in Fig. 1A. Each joint histogram shows values of a pair of pixels with a given relative spatial displacement. Implicit in these measurements is the assumption of homogeneity mentioned in the introduction: The distributions are assumed to be independent of the absolute location within the image. And, although the pixels were taken from a single photographic image (in this case, a New York City street scene), they are nevertheless representative of what one sees in most visual images.

The most striking behavior observed in the plots is that the pixel values are highly correlated: When one is large, the other tends to also be large. But this correlation falls with the distance between pixels. This behavior is summarized in Fig. 1B, which shows the image autocorrelation (pixel correlation as a function of separation).

The correlation statistics of Fig. 1 place a strong constraint on the structure of images, but they do not provide a full probability model. Specifically, there are many probability densities that would share the same correlation (or equivalently, covariance) structure. How should we choose a model

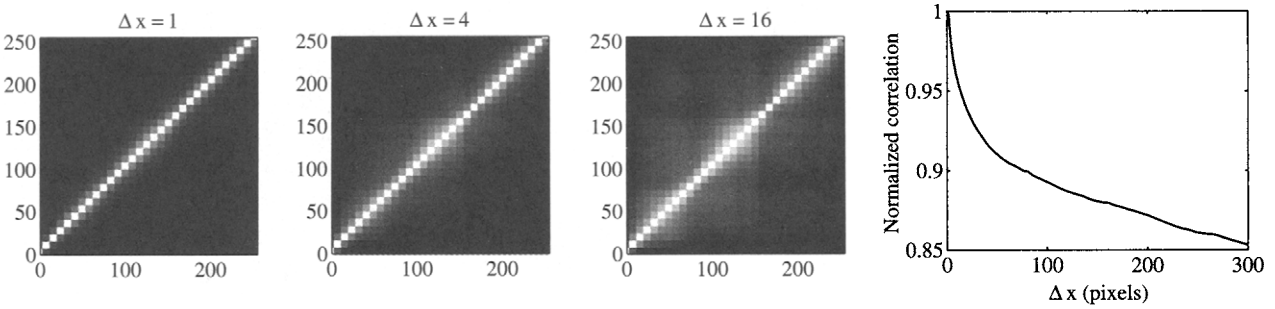

FIGURE 1 **A:** Joint histograms of pairs of pixels at three different spatial displacements, averaged over five examples images. **B:** Autocorrelation function. Photographs are of New York City street scenes, taken with a Canon 10D digital camera, and processed in RAW linear sensor mode (producing pixel intensities are in roughly proportional to light intensity). Correlations were computed on the logs of these sensor intensity values [41].

from this set? One natural solution is to select a density that has maximal entropy, subject to the covariance constraint [25]. Solving for this density turns out to be relatively straighforward, and the result is a multidimensional Gaussian:

$$\mathcal{P}(\vec{x}) \propto \exp(-\vec{x}^T \mathbf{C}_x{}^{-1} \vec{x}/2), \tag{1}$$

where $\vec{x}$ is a vector containing all of the image pixels (assumed, for notational simplicity, to be zero-mean) and $\mathbf{C}_x \equiv \mathsf{E}(\vec{x}\vec{x}^T)$ is the covariance matrix.

Gaussian densities are more succinctly described by transforming to a coordinate system in which the covariance matrix is diagonal. This is easily achieved using standard linear algebra techniques:

$$\vec{y} = E^T \vec{x},$$

where E is an orthogonal matrix containing the eigenvectors of $\mathbf{C}_x$, such that

$$\mathbf{C}_x = EDE^T, \qquad \Longrightarrow E^T \mathbf{C}_x E = D, \tag{2}$$

with D a diagonal matrix containing the associated eigenvalues. When the probability distribution on $\vec{x}$ is stationary (assuming periodic handling of boundaries), the covariance matrix, $\mathbf{C}_x$, will be *circulant*. In this special case, the Fourier transform is known in advance to be a diagonalizing transformation matrix E, and is guaranteed to satisfy the relationship of Equation (2).

To complete the Gaussian image model, we need only specify the entries of the diagonal matrix D, which correspond to the variances of frequency components in the Fourier transform. There are two means of arriving at an answer. First, setting aside the caveats mentioned in the introduction, we can assume that image statistics are scale-invariant. Specifically, suppose that the second-order (covariance) statistical properties of the image are invariant to resizing of the image. We can express scale-invariance in the Fourier domain as:

$$\mathsf{E}(|F(s\omega)|^2) = h(s)\mathsf{E}(|F(\omega)|^2), \qquad \forall \omega, s.$$

That is, rescaling the frequency axis does not change the shape of the function; it merely multiplies the spectrum by a constant. The only functions that satisfy this identity are power laws:

$$\mathsf{E}(|F(\omega)|^2) = \frac{A}{\omega^\gamma}$$

where the exponent γ controls the rate at which the spectrum falls. Thus, the dual assumptions of translation- and scale-invariance constrains the covariance structure of images to a model with only two parameters!

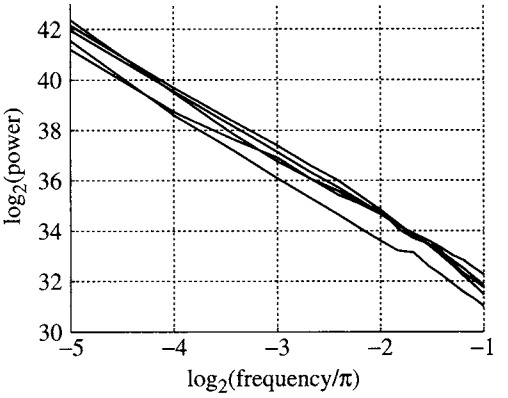

FIGURE 2 Power spectral estimates for five example images (see Fig. 1 for image description), as a function of spatial frequency, averaged over orientation. These are well described by power law functions with an exponent, γ, slightly larger than 2.0.

Alternatively, the form of the power spectrum may be estimated empirically [e.g., 14, 18, 42, 50, 53]. For many "typical" images, it turns out to be quite well approximated by a power law, thus providing confirmation of the scale-invariance property for second-order statistics. In these empirical measurements, the value of the exponent is typically near two. Examples of power spectral estimates for several example images are shown in Fig. 2. It has also been demonstrated that scale-invariance holds for statistics other than the power spectrum [e.g., 42, 52].

The spectral model is *the* classic model of image processing. In addition to accounting for spectra of typical image data, the simplicity of the Gaussian form leads to direct solutions for image compression and denoising that may be found in essentially any textbook on image processing. As an example, consider the problem of removing additive Gaussian white noise from an image, $\vec{x}$. The degradation process is described by the conditional density of the observed (noisy) image, $\vec{y}$, given the original (clean) image $\vec{x}$:

$$\mathcal{P}(\vec{y}|\vec{x}) \propto \exp(-||\vec{y} - \vec{x}||^2/2\sigma_n^2)$$

where σ_n^2 is the variance of the noise. Using Bayes' rule, we can reverse the conditioning by multiplying by the prior probability density on $\vec{x}$:

$$\mathcal{P}(\vec{x}|\vec{y}) \propto \exp(-||\vec{y} - \vec{x}||^2/2\sigma_n^2) \cdot \mathcal{P}(\vec{x}).$$

An estimate for $\vec{x}$ may now be obtained from this posterior density. One can, for example, choose the $\vec{x}$ that maximizes the probability (the *maximum a posteriori* or MAP estimate), or the mean of the density (the *Bayes least squares* or BLS estimate). In the case of a Gaussian prior of Equation (1), these two solutions are identical.

$$\hat{x}(\vec{y}) = \mathbf{C}_x(\mathbf{C}_x + \sigma_n^2 I)^{-1}\vec{y}.$$

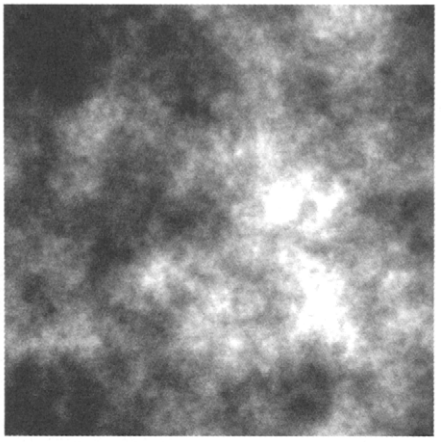

FIGURE 3 Example image randomly drawn from the Gaussian spectral model, with $\gamma = 2.0$.

The solution is linear in the observed (noisy) image $\vec{y}$. Finally, the solution may be rewritten in the Fourier domain, where the scale-invariance of the power spectrum may be explicitly incorporated:

$$\hat{X}(\omega) = \frac{A/\omega^{\gamma}}{A/\omega^{\gamma} + \sigma_n^2} \cdot Y(\omega),$$

where $\hat{X}(\omega)$ and $Y(\omega)$ are the Fourier transforms of $\hat{x}(\vec{y})$ and $\vec{y}$, respectively. Thus, the estimate may be computed by linearly rescaling each Fourier coefficient. To apply this denoising method, one must be given (or must estimate) the parameters, A, γ, and σ_n.

Despite the simplicity and tractability of the Gaussian model, it is easy to see that the model provides a rather weak description. In particular, while the model strongly constrains the amplitudes of the Fourier coefficients, it places no constraint on their *phases*. When one randomizes the phases of an image, the appearance is completely destroyed [37].

As a direct test, one can draw sample images from the distribution by simply generating white noise in the Fourier domain, weighting each sample appropriately by $1/\omega^{\gamma}$, and then inverting the transform to generate an image. An example is shown in Fig. 3. The fact that such an experiment invariably produces images of clouds implies that a covariance constraint is insufficient to capture the richer structure of features that are found in most real images.

3 Wavelet Marginal Models

For decades, the inadequacy of the Gaussian model was apparent. But direct improvement, through introduction of constraints on the Fourier phases, turned out to be quite difficult. Relationships between phase components are not easily measured, in part because of the difficulty of working with joint statistics of circular variables and in part because the dependencies between phases of different frequencies do not seem to be well captured by a model that is localized in frequency. A breakthrough occurred in the 1980s, when a number of authors began to report more direct indications of non-Gaussian behaviors in images. A multidimensional Gaussian statistical model has the property that all conditional or marginal densities must also be Gaussian. But these authors noted that histograms of bandpass-filtered natural images were highly non-Gaussian [9, 13, 18, 31, 58]. These marginals tend to be much more sharply peaked at zero, with more extensive tails, when compared with a Gaussian of the same variance. As an example, Fig. 4 shows histograms of three images, filtered with a Gabor function bandpass filter of octave bandwidth. The intuitive reason for this behavior is that images typically contain smooth regions, punctuated by localized "features" such as (lines, edges, or corners). The smooth regions lead to small filter responses that generate the sharp peak at zero, and the localized features produce large-amplitude responses that generate the extensive tails.

This basic behavior holds for essentially any bandpass filter, whether it is nondirectional (center-surround), or oriented, but some filters lead to responses that are more non-Gaussian than others. By the mid 1990s, a number of authors had developed methods of optimizing a basis of filters to maximize the non-Gaussianity of the responses [e.g., 4, 36]. Often these methods operate by optimizing a higher order statistic such as kurtosis (the fourth moment divided by the squared variance).

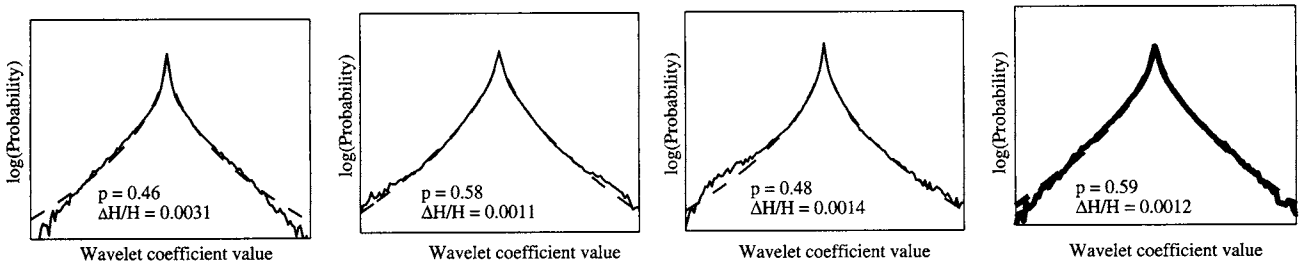

FIGURE 4 Log histograms of a single wavelet subband of four example images (see Fig. 1 for image description). For each histogram, tails are truncated to show 99.8% of the distribution. Also shown (*dashed lines*) are fitted model densities corresponding to equation (3). Text indicates the maximum-likelihood 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.

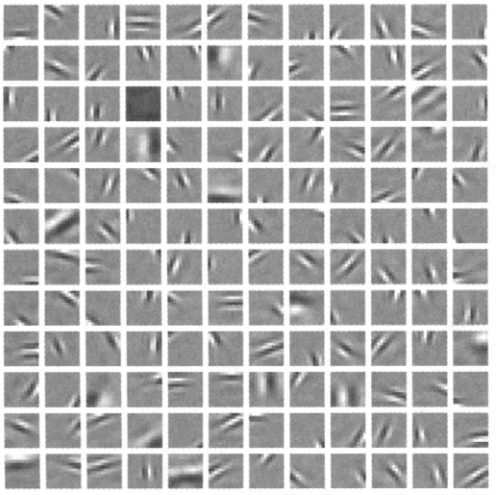

FIGURE 5 Example basis functions derived by optimizing a marginal kurtosis criterion [see 35].

The resulting basis sets contain oriented filters of different sizes with frequency bandwidths of roughly one octave. Figure 5 shows an example basis set, obtained by optimizing kurtosis of the marginal responses to an ensemble of 12×12-pixel blocks drawn from a large ensemble of natural images. In parallel with these statistical developments, authors from a variety of communities were developing multiscale orthonormal bases for signal and image analysis, now generically known as "wavelets" (see Chapter 4.2). These provide a good approximation to optimized bases such as that shown in Fig. 5.

Once we have transformed the image to a multiscale wavelet representation, what statistical model can we use to characterize the the coefficients? The motivation for the choice of basis came from the shape of the marginals, and thus it would seem natural to assume that the coefficients within a subband are independent and identically distributed. With this assumption, the model is completely determined by the marginal statistics of the coefficients, which can be examined empirically as in the examples of Fig. 4. For natural images, these histograms are surprisingly well described by a two-parameter generalized Gaussian (also known as a *stretched,* or *generalized* exponential) distribution [e.g., 31, 34, 47]:

$$\mathcal{P}_c(c; s, p) = \frac{\exp(-|c/s|^p)}{Z(s,p)}, \tag{3}$$

where the normalization constant is $Z(s,p) = 2\frac{s}{p}\Gamma(\frac{1}{p})$. An exponent of $p=2$ corresponds to a Gaussian density, and $p=1$ corresponds to the Laplacian density. In general, smaller values of p lead to a density that is both more concentrated at zero and has more expansive tails. Each of the histograms in Fig. 4 is plotted with 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. The density model fits the histograms remarkably well, as indicated numerically by the relative entropy measures given below each plot. We have observed that values of the exponent p typically lie in the range $[0.4, 0.8]$. The factor s varies monotonically with the scale of the basis functions, with correspondingly higher variance for coarser-scale components.

This wavelet marginal model is significantly more powerful than the classic Gaussian (spectral) model. For example, when applied to the problem of compression, the entropy of the distributions described above is significantly less than that of a Gaussian with the same variance, and this leads directly to gains in coding efficiency. In denoising, the use of this model as a prior density for images yields to significant improvements over the Gaussian model [e.g., 34, 47, 48]. Consider again the problem of removing additive Gaussian white noise from an image. If the wavelet transform is orthogonal, then the noise remains white in the wavelet domain. The degradation process may be described in the wavelet domain as:

$$\mathcal{P}(d|c) \propto \exp(-(d-c)^2/2\sigma_n^2)$$

where d is a wavelet coefficient of the observed (noisy) image, c is the corresponding wavelet coefficient of the original (clean) image, and σ_n^2 is the variance of the noise. Again, using Bayes rule, we can reverse the conditioning:

$$\mathcal{P}(c|d) \propto \exp(-(d-c)^2/2\sigma_n^2) \cdot \mathcal{P}(c),$$

where the prior on c is given by Equation (3). The MAP and BLS solutions cannot, in general, be written in closed form, but numeric solutions are fairly easy to compute [2, 11, 34, 47, 48, and 64]. The resulting estimators are nonlinear "coring" functions, in which small-amplitude coefficients are suppressed and large-amplitude coefficients preserved. These estimates show substantial improvement over the linear estimates associated with the Gaussian model of the previous section (see examples in Fig. 10).

Despite these successes, it is again easy to see that important attributes of images are not captured by wavelet marginal models. When the wavelet transform is orthonormal, one can easily draw statistical samples from the model. Figure 6 shows the result of drawing the coefficients of a wavelet representation independently from generalized Gaussian 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, and perhaps more than the image drawn from the spectral model (Fig. 3), the result still does not look very much like a photographic image!

The wavelet marginal model may be improved by extending it to an *overcomplete* wavelet basis. In particular, Zhu et al. [62] have pointed out, using a variant of the Fourier

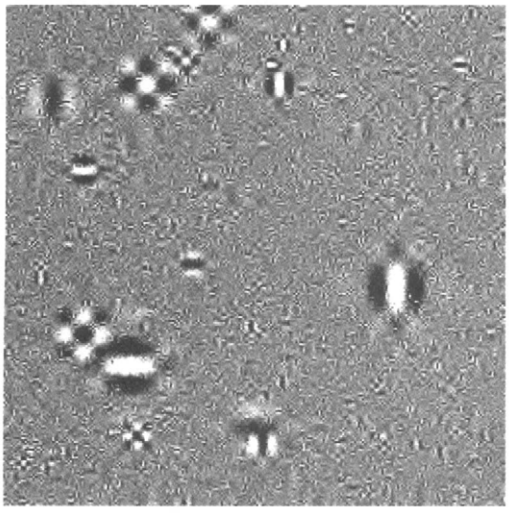

FIGURE 6 A sample image drawn from the wavelet marginal model, with subband density parameters chosen to fit the image of Fig. 7.

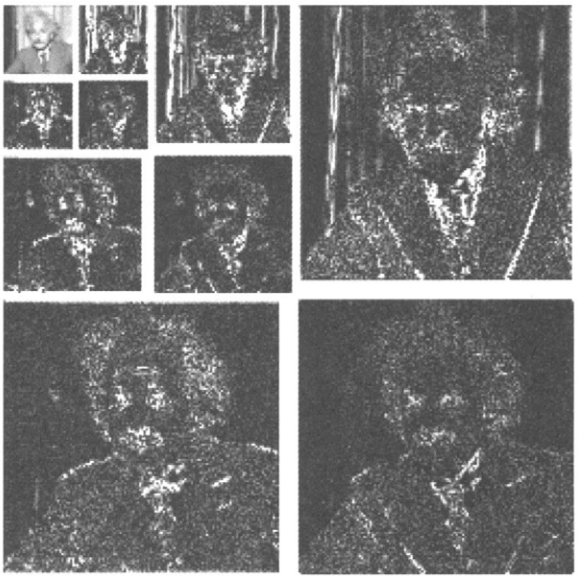

FIGURE 7 Amplitudes of multiscale wavelet coefficients for the "Einstein" image. Each subimage shows coefficient amplitudes of a subband obtained by convolution with a filter of a different scale and orientation, and subsampled by an appropriate factor. Coefficients that are spatially near each other within a band tend to have similar amplitudes. In addition, coefficients at different orientations or scales but in nearby (relative) spatial positions tend to have similar amplitudes.

projection-slice theorem used for tomographic reconstruction, that large numbers of marginals are sufficient to uniquely constrain a high-dimensional probability density. This idea has been used to construct effective models of texture representation and synthesis [20, 61]. The drawback of this approach is that the joint statistical properties are defined *implicitly* through the imposition of marginal statistics. They are thus difficult to study directly or to use in developing solutions for image processing applications. In the next section, we consider the more direct development of joint statistical descriptions.

4 Wavelet Joint Models

The primary reason for the poor appearance of the image in Fig. 6 is that the coefficients of the wavelet transform are not independent. Empirically, the coefficients of orthonormal wavelet decompositions of visual images are found to be nearly decorrelated (i.e., their covariance is zero). But this is only a statement about their *second-order* dependence, and one can easily see that there are important higher-order dependencies. Figure 7 shows the amplitudes (absolute values) of coefficients in a four-level separable orthonormal 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., 7, 46].

The intuitive reason for the clustering of large-amplitude coefficients is that typical localized and isolated image features are represented in the wavelet domain via the superposition of a group of basis functions at different positions, orientations and 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 underlying feature. The 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.

This clustering property was exploited in a heuristic but highly effective manner in the embedded zerotree wavelet (EZW) image coder [44] and has been used in some fashion in nearly all image compression systems since. A more explicit description had been first developed in the context of denoising. More than 20 years ago, Lee [28] suggested a two-step procedure for image denoising in which the local signal variance is first estimated from a neighborhood of observed pixels, after which the pixels in the neighborhood are denoised using a standard linear least squares method. Although it was done in the pixel domain, Lee's paper introduced the idea that variance is a local property that should be estimated *adaptively,* as compared with the classical Gaussian model in which one assumes a fixed global variance. Ruderman [41] examined local variance properties of image derivatives and noted that the derivative field could be made more homogeneous by normalizing by a local estimate of the standard deviation. It was not until the 1990s that a number of authors began to apply this concept to denoising in the wavelet domain, estimating the variance of clusters of wavelet coefficients at nearby positions, scales, and/or orientations, and then using these estimated variances to denoise the cluster [1, 10, 30, 33, 46, 47, 55].

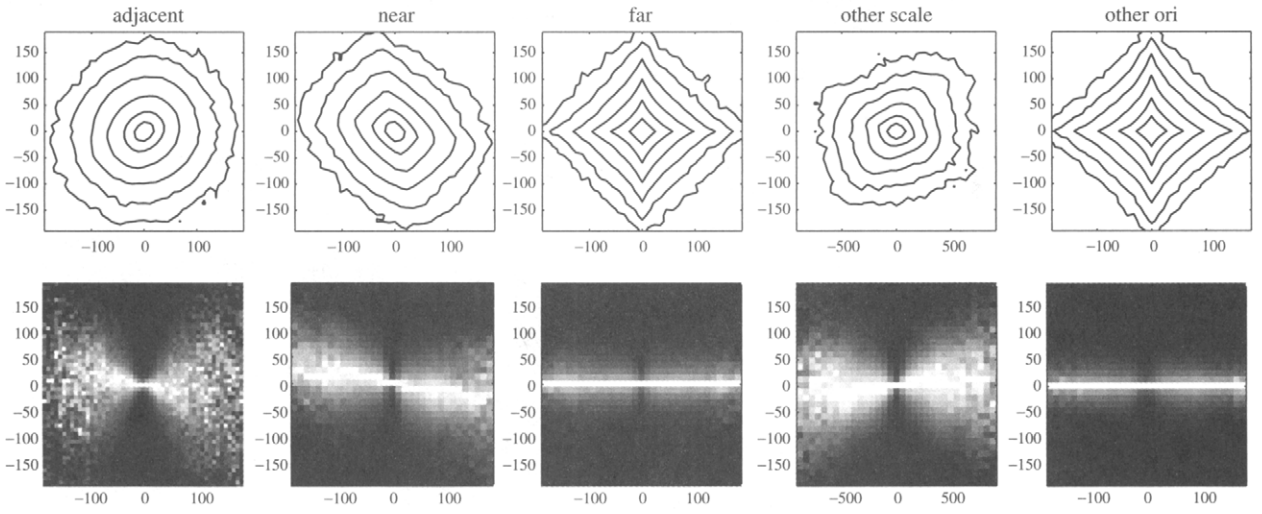

FIGURE 8 Empiric joint distributions of wavelet coefficients associated with different pairs of basis functions, for a single image of a New York City street scene (see Fig. 1 for image description). The top row shows joint distributions as contour plots, with lines drawn at equal intervals of log probability. The three left-most examples correspond to pairs of basis functions at the same scale and orientation, but separated by different spatial offsets. The next corresponds to a pair at adjacent scales (but the same orientation, and nearly the same position), and the right-most example corresponds to a pair at orthogonal orientations (but the same scale and nearly the same position). The bottom row shows corresponding conditional distributions: brightness corresponds to frequency of occurance, except that each column has been independently rescaled to fill the full range of intensities.

The locally adaptive variance principle is powerful, but does not constitute a full probability model. As in the previous sections, we can develop a more explicit model by directly examining the statistics of the coefficients [46]. The top row of Fig. 8 shows joint histograms of several different pairs of wavelet coefficients. As with the marginals, we assume homogeneity to consider the joint histogram of this pair of coefficients, gathered over the spatial extent of the image, as re-presentative of the underlying density. Coefficients that come from adjacent basis functions are seen to produce contours that are nearly circular, whereas the others are clearly extended along the axes. Zetzsche et al. [59] has examined the empiric joint densities of quadrature (Hilbert transform) pairs of basis functions and found that the contours are roughly circular. Several authors have also suggested circular generalized Gaussians as a model for joint statistics of nearby wavelet coefficients [22, 49].

The joint histograms shown in the top row of Fig. 8 do not make explicit the issue of whether the coefficients are independent. The bottom row shows *conditional* histograms of the same data. Let x_2 correspond to the density coefficient (vertical axis), and x_1 the conditioning coefficient (horizontal axis). The histograms illustrate several important aspects of the relationship between the two coefficients. First, apart from the "near" pair, the expected value of x_2 is approximately zero for all values of x_1, indicating that they are nearly decorrelated (to second order). Second, the variance of the conditional histogram of x_2 clearly depends on the value of x_1, and the strength of this dependency depends on the particular pair of coefficients being considered. Thus, even when x_2 and x_1 are uncorrelated, they still exhibit statistical dependence.

The variance dependance shown in Fig. 8 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 and adjacent orientations. As one considers coefficients that are more distant (either in spatial position or in scale), the dependency becomes weaker, suggesting that a Markov assumption might be appropriate.

The circular (or elliptical) contours, the dependency between local coefficient amplitudes, and the associated marginal behaviors, can be modeled using a random field with a spatially fluctuating variance. A particularly useful example arises from the product of a Gaussian vector and a hidden scalar multiplier, known as a *Gaussian scale mixture* (GSM) [3]. These distributions represent an important subset of the *elliptically symmetric distributions,* which are those that can be defined as functions of a quadratic norm of the random vector. Embedded in a random field, these kinds of models have been found useful in the speech-processing community [6]. A related set of models, known as autoregressive conditional heteroskedastic (ARCH) models [e.g., 5], have proven useful for many real signals that suffer from abrupt fluctuations, followed by relative "calm" periods (stock market prices, for example). Finally, physicists studying properties of turbulence have noted similar behaviors [e.g., 51].

Formally, a random vector $\vec{x}$ is a Gaussian scale mixture [3] if and only if it can be expressed as the product of a zero-mean

Gaussian vector $\vec{u}$ and an independent positive scalar random variable $\sqrt{z}$:

$$\vec{x} \sim \sqrt{z}\vec{u}, \tag{4}$$

where $\sim$ indicates equality in distribution. The variable z is known as the *multiplier*. The vector $\vec{x}$ is thus an infinite mixture of Gaussian vectors, whose density is determined by the covariance matrix $\mathbf{C}_u$ of vector $\vec{u}$ and the mixing density, $p_z(z)$:

$$\begin{aligned} p_{\vec{x}}(\vec{x}) &= \int p(\vec{x}|z)\, p_z(z) dz \\ &= \int \frac{\exp\left(-\vec{x}^T (z\mathbf{C}_u)^{-1} \vec{x}/2\right)}{(2\pi)^{N/2} |z\mathbf{C}_u|^{1/2}} p_z(z) dz, \end{aligned} \tag{5}$$

where N is the dimensionality of $\vec{x}$ and $\vec{u}$ (in our case, the size of the neighborhood).

The conditions under which a random vector may be represented using a GSM have been studied [3], and the GSM family includes the α-stable family (including the Cauchy distribution), the generalized Gaussian (or stretched exponential) family, and the symmetrized Gamma family [55]. GSM densities are symmetric and zero-mean, and they have highly kurtotic marginal densities (i.e., heavier tails than a Gaussian). A key property of the GSM model is that the density of $\vec{x}$ is Gaussian when conditioned on z. Also, the normalized vector $\vec{x}/\sqrt{z}$ is Gaussian.

A number of recent image models describe the wavelet coefficients within each local neighborhood using a GSM model, which can capture the strongly leptokurtotic behavior of the marginal densities of natural image wavelet coefficients and the correlation in their local amplitudes, as illustrated in Fig. 9. For example, Baraniuk and colleagues [12, 40] used a two-state hidden multiplier variable to characterize the two modes of behavior corresponding to smooth or low-contrast textured regions and features. Others assume that the local variance is governed by a continuous multiplier variable [29, 33, 39, 54, 55]. Some GSM models for images treat the multiplier variables, z, as if they were independent, even when they belong to overlapping coefficient neighborhoods [29, 39, 54]. More sophisticated models describe dependencies between these variables [12, 40, 55].

The underlying Gaussian structure of the GSM model allows it to be adapted for problems such as denoising. The estimator is more complex than that described for the Gaussian or wavelet marginal models (see [39] for details), but the denoising performance shows a substantial improvement across a wide variety of images and noise levels. As a demonstration, Fig. 10 shows a performance comparison of BLS estimators based on the GSM model, a wavelet marginal model, and a wavelet Gaussian model. The GSM estimator is significantly better, both visually and in terms of mean squared error.

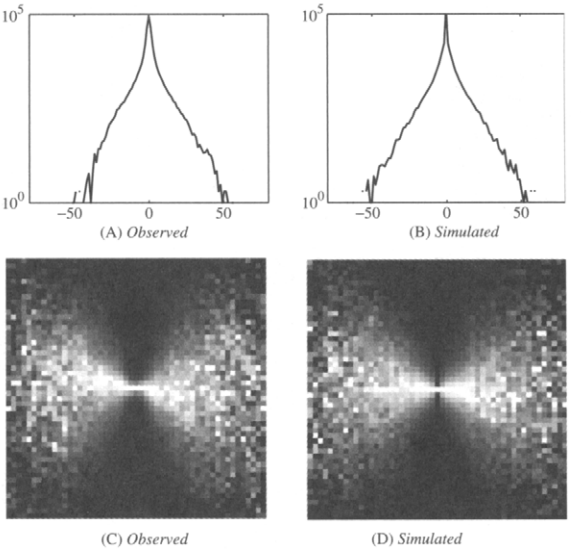

FIGURE 9 Comparison of statistics of coefficients from an example image subband (**left panels**) with those generated by simulation of a local Gaussian scale mixture (GSM) model (**right panels**). Model parameters (covariance matrix and the multiplier prior density) are estimated by maximizing the likelihood of the subband coefficients (see [39]). **A, B:** Log of marginal histograms. **C, D:** Conditional histograms of two spatially adjacent coefficients. Pixel intensity corresponds to frequency of occurance, except that each column has been independently rescaled to fill the full range of intensities.

As with the models of the previous two sections, there are indications that the GSM model is insufficient to fully capture the structure of typical visual images. To demonstrate this, we note that normalizing each coefficient by (the square root of) its estimated variance should produce a field of Gaussian white noise [54]. Figure 11 illustrates this process, showing an example wavelet subband, the estimated variance field, and the normalized coefficients. There are two important types of structure that remain. First, although the normalized coefficients are certainly closer to a homogeneous field, the *signs* of the coefficients still exhibit important structure. Second, the variance field itself is far from homogeneous, with most of the significant values concentrated on one-dimensional contours.

5 Discussion

After nearly 50 years of Fourier/Gaussian modeling, the late 1980s and 1990s saw sudden and remarkable shift in viewpoint, arising from the confluence of (a) multiscale image decompositions, (b) non-Gaussian statistical observations and descriptions, and (c) variance-adaptive statistical models based on hidden variables. The improvements in image processing applications arising from these ideas have been steady and substantial. But the complete synthesis of these ideas, and development of further refinements are still underway.

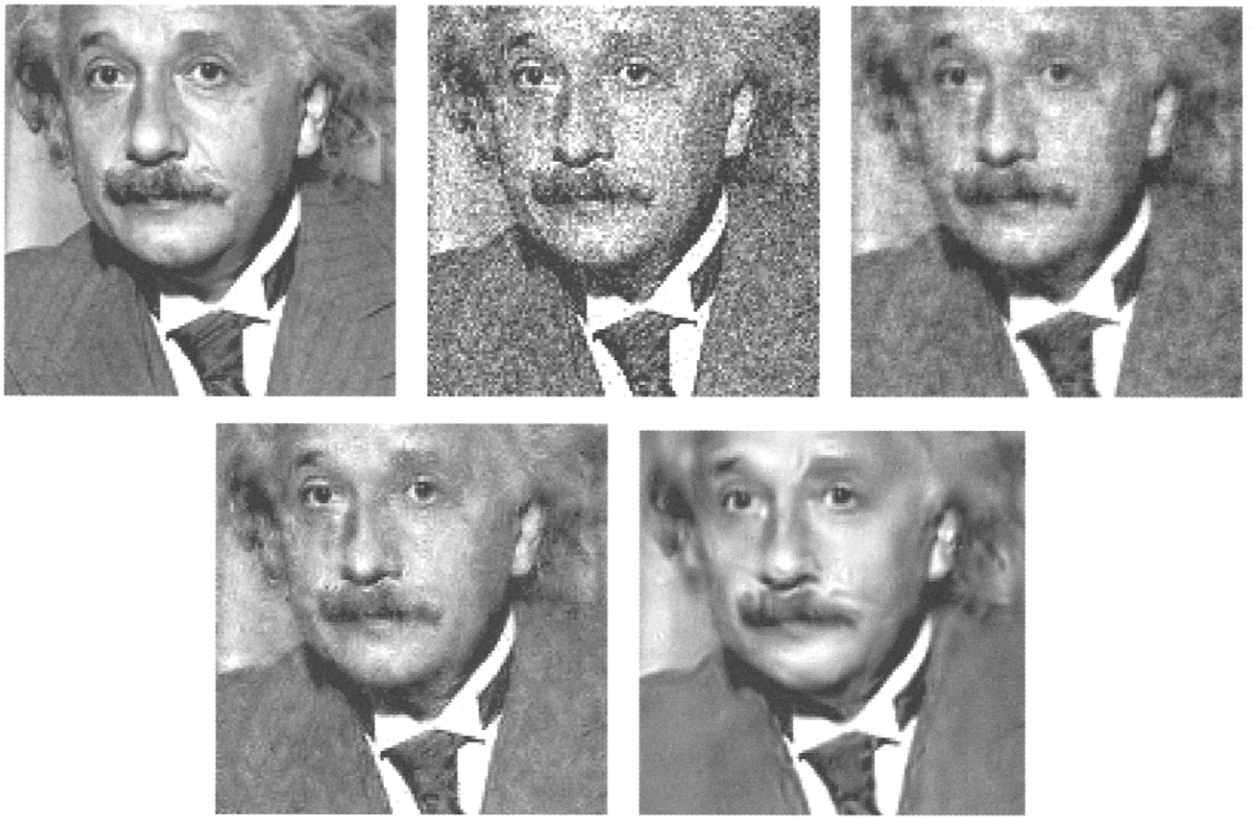

FIGURE 10 Denoising examples. **Upper left:** Original 8-bit image (cropped for visibility). **Upper middle:** Contaminated by adding simulated Gaussian white noise, $\sigma = 21.4$, peak signal-to-noise ratio (PSNR) = 22.06. **Upper right:** Denoised with a Gaussian marginal model (PSNR = 27.87). **Lower left:** Denoised with generalized-Gaussian marginal model (PSNR = 29.24). **Lower right:** Denoised with a Gaussian scale mixture (GSM) model (PSNR = 30.86). All methods were implemented in an overcomplete wavelet domain (see [39, 47]). In each case, the noise variance was assumed known, and the denoising procedure was Bayes least-squares (i.e., the mean of the posterior distribution). Model parameters were estimated by maximizing the likelihood of the data.

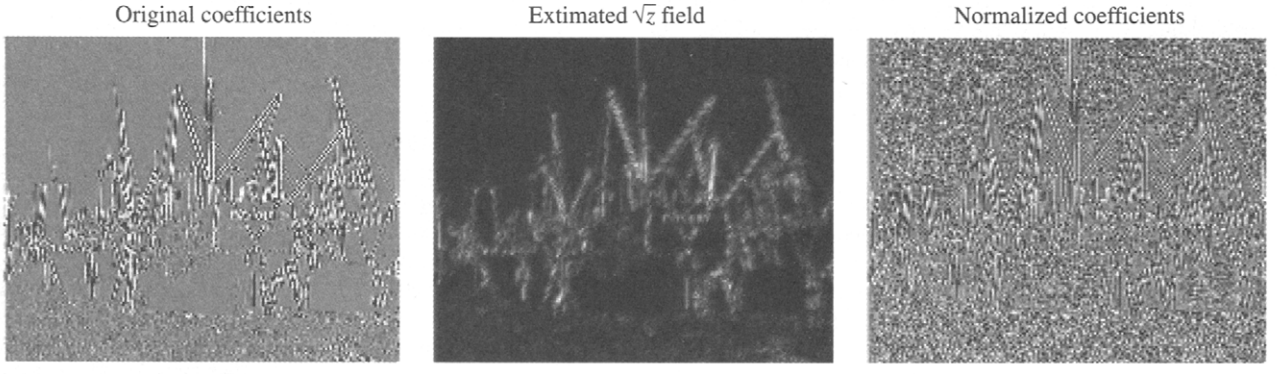

FIGURE 11 Example wavelet subband, square root of the variance field, and normalized subband.

Variants of the GSM model described in the previous section seem to represent the current state of the art, both in terms of characterizing the density of coefficients and in terms of the quality of results in image processing applications. There are several issues that seem to be of primary importance in trying to extend such models. First, a number of authors have examined different methods of describing regularities in the local variance field. These include spatial random fields [23, 24, 26], and multiscale tree-structured models [40, 55]. Much of the structure in the variance field may be attributed to discontinuous features such as edges, lines, or corners. There is a substantial literature in computer vision describing such structures [e.g., 17, 27, 32, 56, 57], but they have proven difficult to establish statistical models that are both explicit

and flexible. Finally, there have been several recent studies investigating geometric regularities that arise from the continuity of contours and boundaries [16, 19, 21, 45, 60, 63]. These and other image structures will undoubtedly be incorporated into future statistical models, leading to further improvements in image processing applications.

References

[1] F. Abramovich, T. Besbeas, and T. Sapatinas, "Empirical Bayes approach to block wavelet function estimation," *Comput. Stat. Data Anal.* 39, 435–451 (2002).

[2] F. Abramovich, T. Sapatinas, and B. W. Silverman, "Wavelet thresholding via a Bayesian approach," *J R Stat Soc B* 60, 725–749 (1998).

[3] D. Andrews and C. Mallows, "Scale mixtures of normal distributions," *J. Royal Stat. Soc.* 36, 99–102 (1974).

[4] A. J. Bell and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision Res.* 37, 3327–3338 (1997).

[5] T. Bollersley, K. Engle, and D. Nelson, "ARCH models," in B. Engle and D. McFadden, ed, *Handbook of Econometrics.* 4, 49, 2959–3038, Elsevier, 1999.

[6] H. Brehm and W. Stammler, "Description and generation of spherically invariant speech-model signals," *Signal Process.* 12, 119–141 (1987).

[7] R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Trans Image Proc* 8, 1688–1701 (1999).

[8] G. Buchsbaum and A. Gottschalk, "Trichromacy, opponent color coding, and optimum colour information transmission in the retina," *Proc. R. Soc. Lond. Ser.* B 220, 89–113 (1983).

[9] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Comm.* COM-31, 532–540 (1983).

[10] S. G. Chang, B. Yu, and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," In *Fifth IEEE Int'l Conf on Image Proc* (Chicago, October 1998).

[11] H. A. Chipman, E. D. Kolaczyk, and R. M. McCulloch, "Adaptive Bayesian wavelet shrinkage," *J. Am. Statist. Assoc.* 92, 1413–1421 (1997).

[12] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Proc.* 46, 886–902 (1998).

[13] J. G. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression, *IEEE Trans. Acoust. Speech Signal Proc.* 36, 1169–1179 (1988).

[14] N. G. Deriugin, "The power spectrum and the correlation function of the television signal," *Telecommunications* 1, 1–12 (1956).

[15] D. W. Dong and J. J. Atick, "Statistics of natural time-varying images," *Network: Comput. Neural Syst.* 6, 345–358 (1995).

[16] J. H. Elder and R. M. Goldberg, "Ecological statistics of gestalt laws for the perceptual organization of contours," *J. Vision* 2, 324–353 (2002).

[17] J. H. Elder and S. W. Zucker, "Local scale control for edge detection and blur estimation," *IEEE Pat. Anal. Mach. Intell.* 20, 699–716 (1998).

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

[19] W. S. Geisler, J. S. Perry, B. J. Super *et al.*, "Edge co-occurance in natural images predicts contour grouping performance," *Vision Res.* 41, 711–724 (2001).

[20] D. Heeger and J. Bergen, "Pyramid-based texture analysis/ synthesis," In *Proc. ACM SIGGRAPH*, 229–238 (1995).

[21] P. Hoyer and A. Hyvärinen, "A multi-layer sparse coding network learns contour coding from natural images," *Vision Res.* 42, 1593–1605 (2002).

[22] J. Huang and D. Mumford, "Statistics of natural images and models," *CVPR* 216 (1999).

[23] A. Hyvärinen and P. Hoyer, "Emergence of topography and complex cell properties from natural images using extensions of ICA," in S. A. Solla, T. K. Leen, and K.-R. Müller, ed, *Adv. Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2000) 827–833.

[24] A. Hyvärinen, J. Hurri, and J. Väyrynen, "Bubbles: a unifying framework for low-level statistical properties of natural image sequences," *J. Opt. Soc. Am.* 20, 1237–1252 (2003).

[25] E. T. Jaynes, "Where do we stand on maximum entropy?" in R. D. Levine and M. Tribus, ed, *The Maximal Entropy Formalism* (MIT Press, Cambridge, MA, 1978).

[26] Y. Karklin and M. S. Lewicki, "Learning higher-order structures in natural images," *Network* 14, 483–499 (2003).

[27] P. Kovesi, "Image features from phase congruency," *Videre: J. Comput. Vis. Res.* 1, 3 (1999).

[28] J. S. Lee, "Digital image enhancement and noise filtering by use of local statistics," *IEEE Pat. Anal. Mach. Intell.* PAMI-2, 165–168 (1980).

[29] 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, 1997).

[30] M. Malfait and D. Roose, "Wavelet-based image denoising using a Markov random field a priori model," *IEEE Trans. Image Proc.* 6, 549–565 (1997).

[31] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Pat. Anal. Mach. Intell.* 11, 674–693 (1989).

[32] S. G. Mallat, "Zero-crossings of a wavelet transform. *IEEE Trans. Info. Theory* 37, 1019–1033 (1991).

[33] M. K. Mihçak, I. Kozintsev, K. Ramchandran, *et al.*, "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Trans. Sig. Proc.* 6, 300–303 (1999).

[34] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using a generalized Gaussian and complexity priors," *IEEE Trans. Info. Theory* 45, 909–919 (1999).

[35] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature* 381, 607–609 (1996).

[36] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Res.* 37, 3311–3325 (1997).

[37] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proc. IEEE* 69, 529–541 (1981).

[38] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int. J. Comput. Vision* 40, 49–71 (2000).

[39] J. Portilla, V. Strela, M. Wainwright, *et al.*, "Image denoising using a scale mixture of Gaussians in the wavelet domain," *IEEE Trans Image Process.* 12, 1338–1351 (2003).

[40] J. Romberg, H. Choi, and R. Baraniuk, "Bayesian wavelet domain image modeling using hidden Markov trees," in *Proc. IEEE Int. Conf Image Proc* (Kobe, Japan, 1999).

[41] D. L. Ruderman, "The statistics of natural images," *Network: Comput. Neural Syst.* 5, 517–548 (1996).

[42] D. L. Ruderman and W. Bialek, "Statistics of natural images: scaling in the woods," *Phys. Rev. Lett.* 73, 814–817 (1994).

[43] D. L. Ruderman, T. W. Cronin, and C.-C. Chiao, "Statistics of cone responses to natural images: Implications for visual coding," *J. Opt. Soc. Am.* 15, 2036–2045 (1998).

[44] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Sign. Proc.* 41, 3445–3462 (1993).

[45] M. Sigman, G. A. Cecchi, C. D. Gilbert, *et al.*, "On a common circle: natural scenes and Gestalt rules," *Proc. Nat. Acad. Sciences* 98, 1935–1940 (2001).

[46] E. P. Simoncelli, "Statistical models for images: Compression, restoration and synthesis," In *Proc 31st Asilomar Conf on Signals, Systems and Computers* (IEEE Computer Society, Pacific Grove, CA, 1997).

[47] E. P. Simoncelli, "Bayesian denoising of visual images in the wavelet domain. In P. Müller and B. Vidakovic, ed, *Bayesian Inference in Wavelet Based Models* (Springer-Verlag, New York, 1999) 291–308.

[48] E. P. Simoncelli and E. H. Adelson, "Noise removal via Bayesian wavelet coring," in *Third Int'l Conf on Image Proc* I, 379–382 (1996).

[49] A. Srivastava, X. Liu, and U. Grenander, "Universal analytical forms for modeling image probability," *IEEE Pat. Anal. Mach. Intell.* 28, (2002).

[50] D. J. Tolhurst, Y. Tadmor, and T. Chao, "Amplitude spectra of natural images," *Opth. Physiol. Optics* 12, 229–232 (1992).

[51] A. Turiel, G. Mato, N. Parga, et al., "The self-similarity properties of natural images resemble those of turbulent flows," *Phys. Rev. Lett.* 80, 1098–1101 (1998).

[52] A. Turiel and N. Parga, "The multi-fractal structure of contrast changes in natural images: From sharp edges to textures," *Neural Comput.* 12, 763–793 (2000).

[53] A. van der Schaaf and J. H. van Hateren, "Modelling the power spectra of natural images: statistics and information," *Vision Res.* 28, 2759–2770 (1996).

[54] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," in S. A. Solla, T. K. Leen, and K.-R. Müller, ed, *Adv. Neural Information Processing Systems (NIPS'99)* May 2000. (MIT Press, Cambridge, MA, 2000) 855–861.

[55] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in modeling and analyzing natural imagery," *Appl. Comput. Harmonic Anal.* 11, 89–123 (2001).

[56] Z. Wang and E. P. Simoncelli, "Local phase coherence and the perception of blur," in S. Thrun, L. Saul, and B. Schölkopf, ed, *Adv. Neural Information Processing Systems (NIPS*03)*, 2004. (MIT Press, Cambridge, MA, 2004).

[57] A. P. Witkin, "Scale-space filtering," in *Proc. Int. Joint Conf. Artificial Intelligence* 1019–1021 (1985).

[58] C. Zetzsche and E. Barth, "Fundamental limits of linear filters in the visual processing of two-dimensional signals," *Vision Res.* 30, 1111–1117 (1990).

[59] C. Zetzsche, B. Wegmann, and E. Barth, "Nonlinear aspects of primary vision: entropy reduction beyond decorrelation," *in Intl. Symp. Soc. Information Displ.* XXIV, 933–936 (1993).

[60] S.-C. Zhu, "Statistical modeling and conceptualization of visual patterns," *IEEE Trans PAMI* 25, 691–712 (2003).

[61] S. C. Zhu, Y. N. Wu, and D. Mumford, "Minimax entropy principle and its application to texture modeling," *Neural Comput.* 9, 1627–1660 (1997).

[62] S. C. Zhu, Y. N. Wu, and D. Mumford, "FRAME: Filters, random fields and maximum entropy – towards a unified theory for texture modeling," *Int. J. Comp. Vis.* 27, 1–20 (1998).

[63] A. B. Lee, K. S. Pedersen, and D. Mumford, "The Nonlinear Statistics of High-Contrast Patches in Natural Images," *Int' Journal of Computer Vision*, 54, 88–103 (2003).

[64] B. Vidakovic, "Nonlinear wavelet Shrinkage with Bayes rules and Bayes Factors," *J. American Statistical Association*, 93, 173–179 (1998).