9. Capturing Visual Image Properties with Probabilistic Models

Eero P. Simoncelli

New York University

July 18, 2008

The set of all possible visual images is enormous, but not all of these are equally likely to be encountered by your eye or a camera. This nonuniform distribution over the image space is believed to be exploited by biological 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. Image compression amounts 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, subject to some set of constraints. Specific examples of these applications can be found in many chapters throughout this book.

In order 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, etc, (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.

In this chapter, we'll describe an empirical methodology for building and testing probability models for discretized (pixelated) images. Currently available digital cameras record 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 exponentially in D, the dimensionality of the space (in this case, the number of pixels). This is known as the "curse of dimensionality". For example, if we wanted to build a nhistogram for images with one million pixels, and each pixel value was partitioned into just two possibilites (low or high), we'd need $2^{1,000,000}$ bins, which greatly exceeds estimates of the number of atoms in the universe!

Thus, in order to make progress on image modeling, it is *essential* that we 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 (i.e., sometimes called homogeneity, or strict-sense stationarity): the probability density 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 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 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 will also fail for finite size images 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, psychology,applied math and statistics have proposed a wide variety of different types of model. In this chapter, I’ll review the most basic statistical properties of photographic images, and describe several models that have been developed to incorporate these properties. I’ll give some indication of how these models have been validated by examining how well they fit the data. In order to keep the discussion focused, I’ll limit the discussion to discretized *grayscale* photographic images. Many of the principles are easily extended to color photographs [8, 43], or temporal image sequences (movies) [16], as well as more specialized image classes such as portraits, landscapes, or textures. In addition, the general concepts are often applicable to non-visual imaging devices, such as medical images, infrared images, radar and other types of range image, or astronomical images.

1 The Gaussian Model

The classical 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. 1(a). Each scatterplot shows values of a pair of pixels[1] with a different relative horizontal 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.

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. This correlation weakens with the distance between pixels. This behavior is summarized in Fig. 1(b), 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 from amongst this set? One natural criterion is to select a density that has maximal entropy, subject to the covariance constraint [24]. Solving for this density turns out to be relatively straighforward, and the result is a multi-dimensional Gaussian:

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

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

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 [49]:

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

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

$$\mathbf{C_x} = EDE^T, \qquad \Longrightarrow E^T \mathbf{C_x} E = D. \qquad (2)$$

D is 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[2], and is guaranteed to satisfy the relationship of Eq. (2).

[1] Pixel values recorded by digital cameras are generally nonlinearly related to the light intensity that fell on the sensor. Here, we used linear measurements in a single images of a New York City street scene, as recorded by the CMOS sensor, and took the log of these.

[2] More generally, the Fourier transform diagonalizes any matrix that represents a translation-invariant (i.e., convolution) operation.

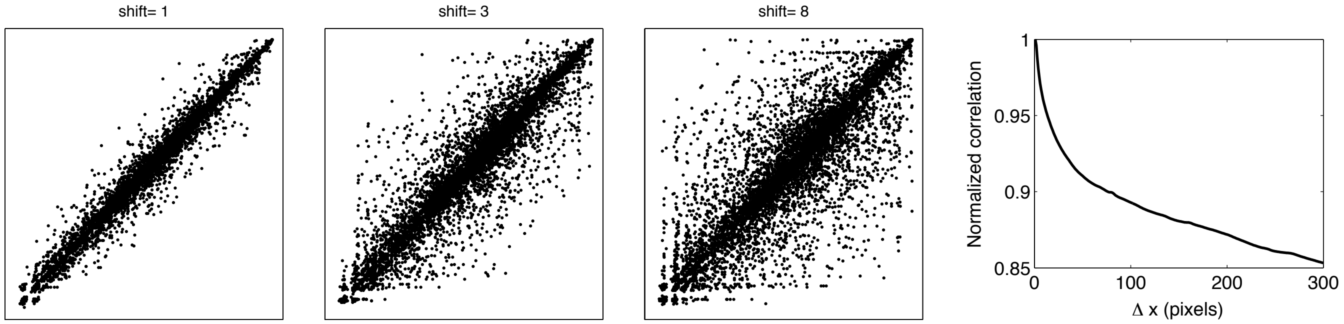


Fig. 1. (a) Scatterplots comparing values of pairs of pixels at three different spatial displacements, averaged over five examples images. (b) Autocorrelation function. Photographs are of New York City street scenes, taken with a Canon 10D digital camera in RAW mode (these are the sensor measurements which are approximately proportional to light intensity). The scatterplots and correlations were computed on the logs of these sensor intensity values [41].

In order 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 frequency domain as:

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

where $F(\vec{\omega})$ indicates the (two-dimensional) Fourier transform of the image. 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:

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

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 two parameters!

Alternatively, the form of the power spectrum may be estimated empirically [e.g. 15, 18, 50, 42, 53]. For many "typical" images, it turns out to be quite well approximated by a power law, consistent with the scale-invariance assumption. 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].

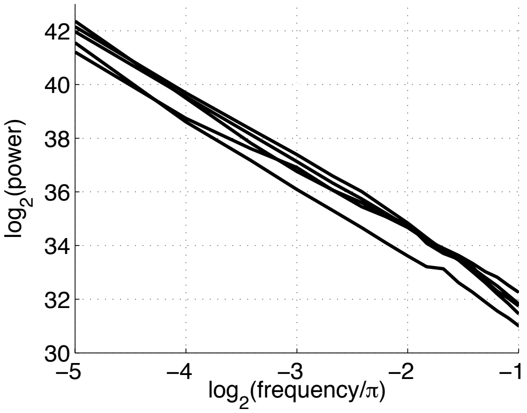


Fig. 2. Power spectral estimates for five example images (see Fig. 1 for image description), as a function of spatial frequency, averaged over orientation. These are well-described by power law functions with an exponent, γ, slightly larger than 2.0.

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 signal or 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 $\vec{x}$:

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

An estimate, $\hat{x}$ 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 *minimum mean squared error* (MMSE) or *Bayes Least Squares* or BLS estimate). If we assume that the prior density is Gaussian, then the posterior density will also be Gaussian, the maximum and the mean will then be identical:

$$\hat{x}(\vec{y}) = \mathbf{C_x}(\mathbf{C_x} + \mathbf{I}\sigma_n^2)^{-1}\vec{y},$$

where $\mathbf{I}$ is an identity matrix. Note that this solution is linear in the observed (noisy) image $\vec{y}$.

This linear estimator is particularly simple when both the noise and signal covariance matrices are diagonalized. As mentioned previously, under the spectral model , the signal covariance matrix may be diagonlized by transforming to the Fourier domain, where the estimator may be written as:

$$\hat{F}(\vec{\omega}) = \frac{A/|\vec{\omega}|^\gamma}{A|\vec{\omega}|^\gamma + \sigma_n^2} \cdot G(\vec{\omega}),$$

where $\hat{F}(\vec{\omega})$ and $G(\vec{\omega})$ 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 individually. In order to apply this denoising method, one must be given (or must estimate) the parameters, A, γ and σ_n (see Chapter 11 for further examples and development of the denoising problem).

Despite the simplicity and tractability of the Gaussian model, it is easy to see that the model provides a rather weak description of images. 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 [36].

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 $1/|\vec{\omega}|^\gamma$, and then inverting the transform to generate an image. The fact that this experiment invariably produces images of clouds (an example is shown in Fig. 3) implies that a Gaussian model is insufficient to capture the structure of features that are found in photographic images.

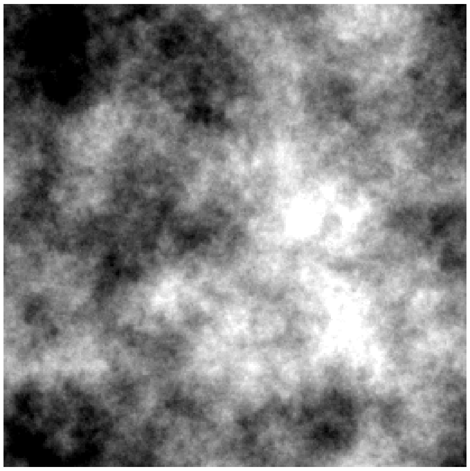


Fig. 3. Example image randomly drawn from the Gaussian spectral model, with $\gamma = 2.0$.

2 The Wavelet Marginal Model

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 describe more direct indications of non-Gaussian behaviors in images. Specifically, 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, 18, 14, 31, 56]. Specifically, their 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 (a Gaussian-windowed sinuosoidal grating). 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 zero-mean local filter, whether it is non-directional (center-surround), or oriented, but some filters lead to responses

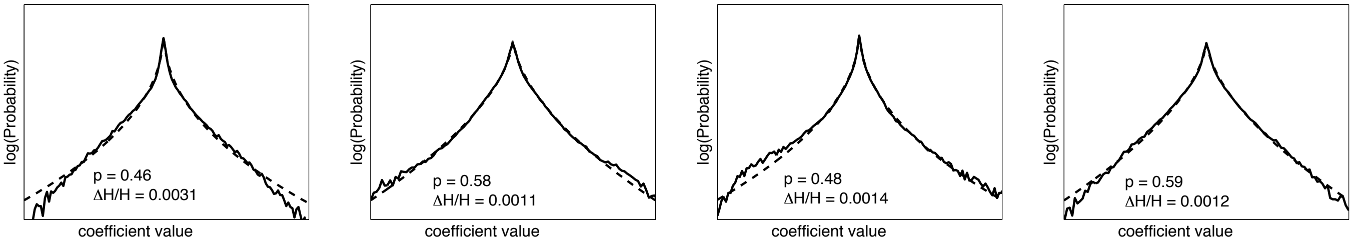


Fig. 4. Log histograms of bandpass (Gabor) filter responses of four example images (see Fig. 1 for image description). For each histogram, tails are truncated so as to show 99.8% of the distribution. Also shown (dashed lines) are fitted generalized Gaussian densities, as specified by equation (3). Text indicates the maximum-likelihood value of p of 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.

that are more non-Gaussian than others. By the mid 1990s, a number of authors had developed methods of optimizing a basis of filters in order to to maximize the non-Gaussianity of the responses [e.g., 35, 5]. Often these methods operate by optimizing a higher-order statistic such as kurtosis (the fourth moment divided by the squared variance). 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 multi-scale orthonormal bases for signal and image analysis, now generically known as "wavelets" (see chapter *** in this book). These provide a good approximation to optimized bases such as that shown in Fig. 5.

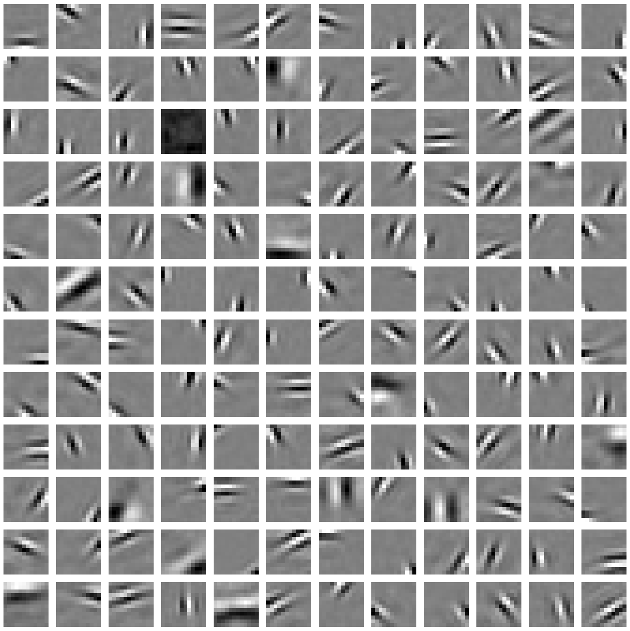


Fig. 5. Example basis functions derived by optimizing a marginal kurtosis criterion [see 34].

Once we've transformed the image to a multi-scale representation, what statistical model can we use to characterize the coefficients? The statistical 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, 47, 33]:

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

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 probability 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 classical 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., 48, 11, 2, 33, 47]. 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:

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

where the prior on c is given by Eq. (3). Here, the MAP and BLS solutions cannot, in general, be written in closed form, and they are unlikely to be the same. But numerical solutions are fairly easy to compute, resulting in nonlinear estimators, 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.

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, we 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 an example photographic 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!

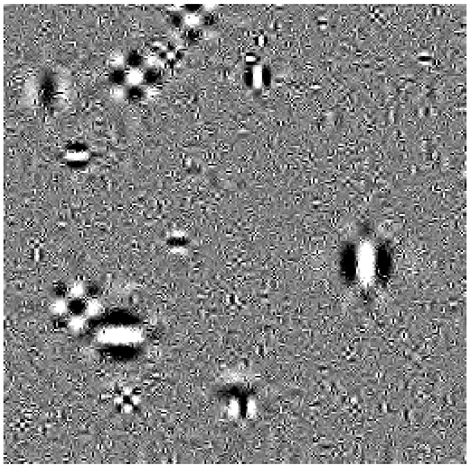


Fig. 6. A sample image drawn from the wavelet marginal model, with subband density parameters chosen to fit the image of Fig. 7.

The wavelet marginal model may be improved by extending it to an *overcomplete* wavelet basis. In particular, Zhu *et al.* have shown that large numbers of marginals are sufficient to uniquely constrain a high-dimensional probability density [58] (this is a variant of the Fourier projection-slice theorem used for tomographic reconstruction). Marginal models have been shown to produce better denoising results when the multi-scale representation is overcomplete [12, 47, 3, 27, 39]. Similar benefits have been obtained for texture representation and synthesis [20, 58]. The drawback of these models is that the joint statistical properties are defined *implicitly* through the marginal statistics. They are thus difficult to study directly, or to utilize in deriving optimal solutions for image processing applications. In the next section, we consider the more direct development of joint statistical descriptions.

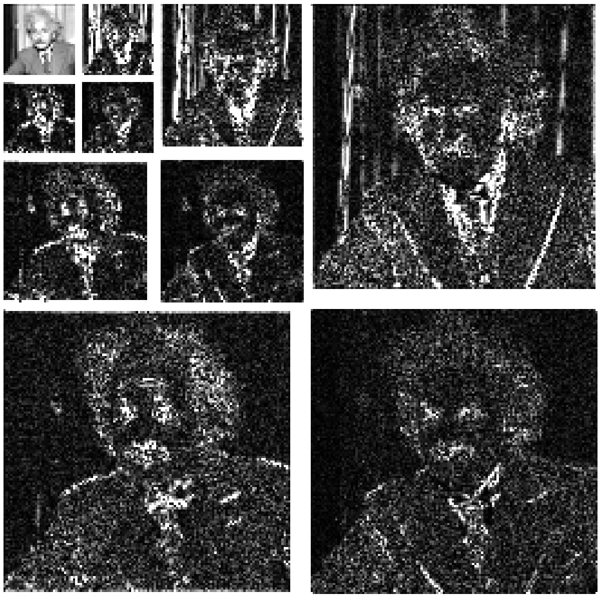


Fig. 7. Amplitudes of multi-scale wavelet coefficients for an image of Albert Einstein. Each subimage shows coefficient amplitudes of a subband obtained by convolution with a filter of a different scale and orientation, and subsampled by an appropriate factor. Coefficients that are spatially near each 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.

3 Wavelet Local Contextual 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 moderately well decorrelated (i.e., their covariance is near 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. First, we can see that individual subbands are not homogeneous: Some regions have large-amplitude coefficients, while other regions are relatively low in amplitude. The variability of the local amplitude is characteristic of most photographic images: The 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.

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 for denoising, when Lee [26] suggested a two-step procedure, 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, this 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. 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 in order to denoise the cluster [30, 46, 10, 47, 32, 55, 1].

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. The top row of Fig. 8 shows joint histograms of several different pairs of wavelet coefficients. As with the marginals, we assume homogeneity 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. 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.

The joint histograms shown in the first row of Fig. 8 do not make explicit the issue of whether the coefficients are independent. In order to make this more explicit, 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, 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, although x_2 and x_1 are uncorrelated, they still exhibit statistical dependence!

The form of the histograms 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.

Essentially all of the statistical properties we've described thus far – the circular (or elliptical) contours, the dependency between local coefficient amplitudes, as well as the heavy-tailed marginals – can be modeled using a random field with a spatially fluctuating variance. These kinds of models have been found useful in the speech-processing community [7]. A related set of models, known as autoregressive conditional heteroskedastic (ARCH) models [e.g., 6], 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].

An example of a local density with fluctuating variance, one that has found particular use in modeling local clusters (neighborhoods) of multi-scale image coefficients, is the product of a Gaussian vector and a hidden scalar multiplier. More formally, this model, known as a *Gaussian scale mixture* [4] (GSM), expresses a random vector $\vec{x}$ 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). Notice that since the level surfaces (contours of constant probability) for $P_{\vec{u}}(\vec{u})$ are ellipses determined by the covariance matrix $\mathbf{C_u}$, and the density of $\vec{x}$ is constructed as a mixture of scaled versions of the density of $\vec{u}$, then $P_{\vec{x}}(\vec{x})$ will *also* exhibit the same elliptical level surfaces. In particular, if $\vec{u}$ is spherically symmetric ($\mathbf{C_u}$ is a multiple of the identity), then $\vec{x}$ will also be spherically symmetric. Figure 9 demonstrates that this model can capture the strongly kurtotic behavior of the marginal densities of natural image wavelet coefficients, as well as the correlation in their local amplitudes.

A number of recent image models describe the wavelet coefficients within each local neighborhood using a Gaussian mixture model [e.g., 13, 40, 28, 32, 55, 38, 29]. Sampling from these models is difficult, since the local description is typically used for *overlapping* neighborhoods, and thus one cannot simply draw independent samples from the model (see [29] for an example). The underlying Gaussian structure of the model allows it to be adapted for problems such as denoising. The resulting estimator is more complex than that described for the Gaussian or wavelet marginal models, but performance is significantly better.

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 [41, 54]. Figure 10 illustrates this process, showing an example wavelet subband, the estimated variance field, and the normalized coefficients. But note that 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. Some of these attributes can be captured by measuring joint statistics of phase and amplitude, as has been demonstrated in texture modeling [37].

4 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) multi-scale

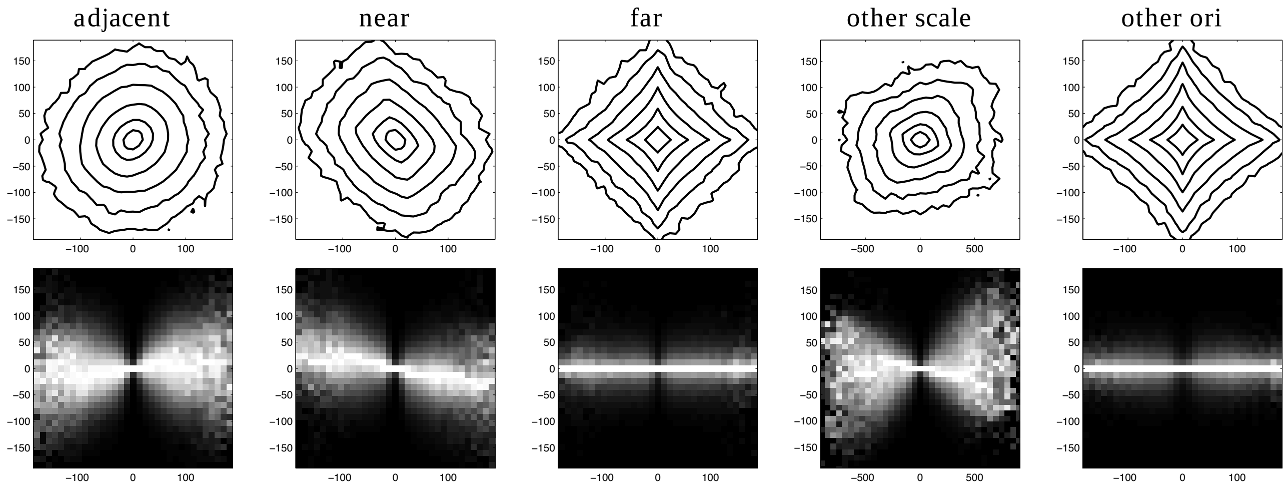

adjacent near far other scale other ori

Fig. 8. Empirical 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 leftmost 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 rightmost 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.

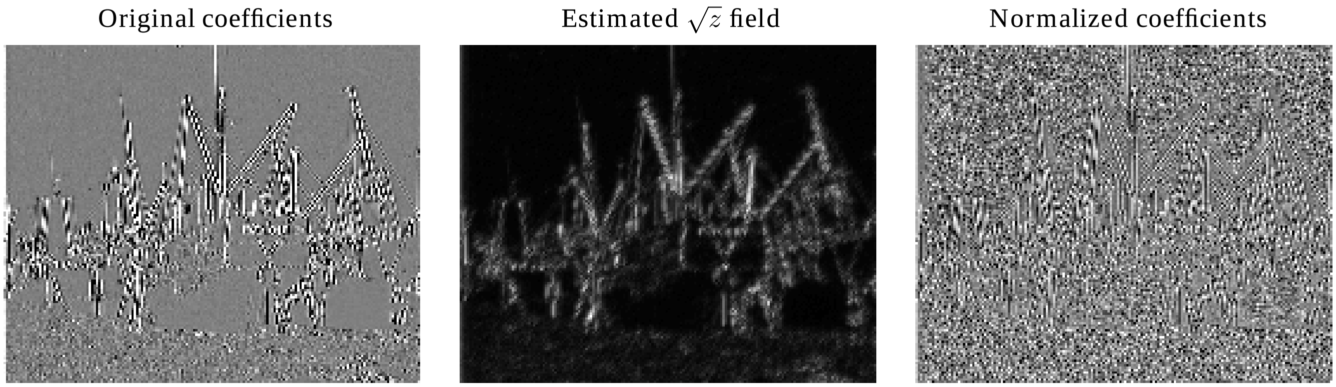
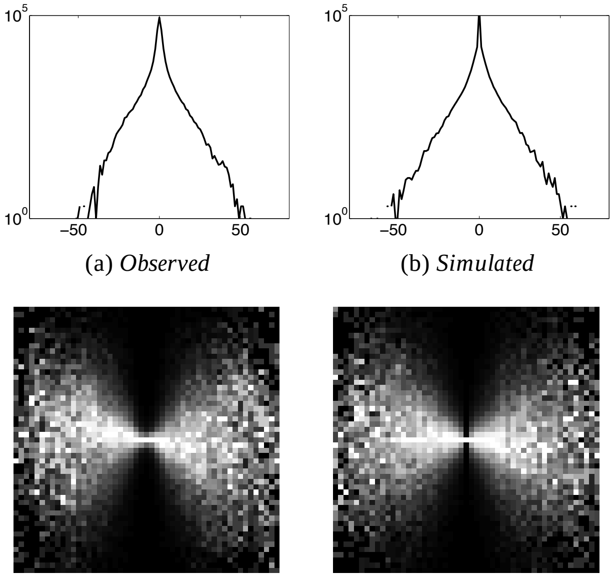

Original coefficients Estimated $\sqrt{z}$ field Normalized coefficients

Fig. 10. Example wavelet subband, square root of the variance field, and normalized subband.



(a) *Observed* (b) *Simulated*

(c) *Observed* (d) *Simulated*

Fig. 9. Comparison of statistics of coefficients from an example image subband (left panels) with those generated by simulation of a local GSM model (right panels). Model parameters (covariance matrix and the multiplier prior density) are estimated by maximizing the likelihood of the subband coefficients (see [38]). **(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.

image decompositions, (b) non-Gaussian statistical observations and descriptions, and (c) locally-adaptive statistical models based on fluctuating variance. 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.

Variants of the contextual models 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 are developing models that can capture the regularities in the local variance, such as spatial random fields [22, 25, 23, 29], 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, but it has proven difficult to establish models that are both explicit about these features and yet flexible. Finally, there have been several recent studies investigating geometric regularities that arise from the continuity of contours and boundaries [45, 17, 19, 21, 57]. These and other image regularities will surely 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. *Computational Statistics and Data Analysis*, 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] F. Abramovich, T. Sapatinas, and B. W. Silverman. Stochastic expansions in an overcomplete wavelet dictionary. *Probability Theory and Related Fields*, 117:133–144, 2000.

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

[5] A. J. Bell and T. J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[6] T. Bollersley, K. Engle, and D. Nelson. ARCH models. In B. Engle and D. McFadden, editors, *Handbook of Econometrics V*. 1994.

[7] H. Brehm and W. Stammler. Description and generation of spherically invariant speech-model signals. *Signal Processing*, 12:119–141, 1987.

[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(4):532–540, April 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. IEEE Computer Society.

[11] H. A. Chipman, E. D. Kolaczyk, and R. M. McCulloch. Adaptive Bayesian wavelet shrinkage. *J American Statistical Assoc*, 92(440):1413–1421, 1997.

[12] R. R. Coifman and D. L. Donoho. Translation-invariant de-noising. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and statistics*. Springer-Verlag lecture notes, San Diego, 1995.

[13] 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, April 1998.

[14] J. G. Daugman. Complete discrete 2–D Gabor transforms by neural networks for image analysis and compression. *IEEE Trans. Acoust. Speech Signal Proc.*, 36(7):1169–1179, 1988.

[15] N. G. Deriugin. The power spectrum and the correlation function of the television signal. *Telecommunications*, 1(7):1–12, 1956.

[16] D. W. Dong and J. J. Atick. Statistics of natural time-varying images. *Network: Computation in Neural Systems*, 6:345–358, 1995.

[17] J. H. Elder and R. M. Goldberg. Ecological statistics of gestalt laws for the perceptual organization of contours. *Journal of Vision*, 2(4):324–353, 2002. DOI 10:1167/2.4.5.

[18] 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, 1987.

[19] W. S. Geisler, J. S. Perry, B. J. Super, and D. P. Gallogly. Edge co-occurance in natural images predicts contour grouping performance. *Vision Research*, 41(6):711–724, March 2001.

[20] D. Heeger and J. Bergen. Pyramid-based texture analysis/synthesis. In *Proc. ACM SIGGRAPH*, pages 229–238. Association for Computing Machinery, August 1995.

[21] P. Hoyer and A. Hyvärinen. A multi-layer sparse coding network learns contour coding from natural images. *Vision Research*, 42(12):1593–1605, 2002.

[22] 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, editors, *Adv. Neural Information Processing Systems*, volume 12, pages 827–833, Cambridge, MA, May 2000. MIT Press.

[23] A. Hyvärinen, J. Hurri, and J. Väyrynen. Bubbles: A unifying framework for low-level statistical properties of natural image sequences. *J. Opt. Soc. Am. A*, 20(7), July 2003.

[24] E. T. Jaynes. Where do we stand on maximum entropy? In R. D. Levine and M. Tribus, editors, *The Maximal Entropy Formalism*. MIT Press, Cambridge, MA, 1978.

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

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

[27] X. Li and M. T. Orchard. Spatially adaptive image denoising under overcomplete expansion. In *IEEE Int'l Conf on Image Proc*, Vancouver, September 2000.

[28] 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, March 1997.

[29] S. Lyu and E. P. Simoncelli. Modeling multiscale subbands of photographic images with fields of Gaussian scale mixtures. *IEEE Trans. Patt. Analysis and Machine Intelligence*, 2008. Accepted for publication, 4/08.

[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, April 1997.

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

[32] M. K. Mihçak, I. Kozintsev, K. Ramchandran, and P. Moulin. Low-complexity image denoising based on statistical modeling of wavelet coefficients. 6(12):300–303, December 1999.

[33] 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.

[34] 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.

[35] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

[36] A. V. Oppenheim and J. S. Lim. The importance of phase in signals. *Proc. of the IEEE*, 69:529–541, 1981.

[37] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *Int'l Journal of Computer Vision*, 40(1):49–71, December 2000.

[38] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli. Image denoising using a scale mixture of Gaussians in the wavelet domain. *IEEE Trans Image Processing*, 12(11):1338–1351, November 2003.

[39] M. Raphan and E. P. Simoncelli. Optimal denoising in redundant representations. *IEEE Trans Image Processing*, 17(8):1342–1352, August 2008.

[40] J. Romberg, H. Choi, and R. Baraniuk. Bayesian wavelet domain image modeling using hidden Markov trees. In *Proc. IEEE Int'l Conf on Image Proc*, Kobe, Japan, October 1999.

[41] D. L. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5:517–548, 1996.

[42] D. L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Phys. Rev. Letters*, 73(6):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. A*, 15(8):2036–2045, 1998.

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

[45] M. Sigman, G. A. Cecchi, C. D. Gilbert, and M. O. Magnasco. On a common circle: Natural scenes and Gestalt rules. *Proc. National Academy of Sciences*, 98(4):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*, volume 1, pages 673–678, Pacific Grove, CA, November 2-5 1997. IEEE Computer Society.

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

[48] E. P. Simoncelli and E. H. Adelson. Noise removal via Bayesian wavelet coring. In *Proc 3rd IEEE Int'l Conf on Image Proc*, volume I, pages 379–382, Lausanne, September 16-19 1996. IEEE Sig Proc Society.

[49] G. Strang. *Linear Algebra and Its Applications*. Academic Press, Orlando, 1980.

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

[51] A. Turiel, G. Mato, N. Parga, and J. P. Nadal. 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 Computation*, 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 Research*, 28(17):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, editors, *Adv. Neural Information Processing Systems (NIPS*99)*, volume 12, pages 855–861, Cambridge, MA, May 2000. MIT Press.

[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. *Applied and Computational Harmonic Analysis*, 11(1):89–123, July 2001.

[56] C. Zetzsche and E. Barth. Fundamental limits of linear filters in the visual processing of two-dimensional signals. *Vision Research*, 30:1111–1117, 1990.

[57] S.-C. Zhu. Statistical modeling and conceptualization of visual patterns. *IEEE Trans PAMI*, 25(6), June 2003.

[58] S. C. Zhu, Y. N. Wu, and D. Mumford. FRAME: Filters, random fields and maximum entropy – towards a unified theory for texture modeling. *Intl. J. Comp. Vis.*, 27(2):1–20, 1998.