Modeling Multiscale Subbands of Photographic Images with Fields of Gaussian Scale Mixtures

Siwei Lyu, *Member, IEEE,* and Eero P. Simoncelli, *Fellow, IEEE*

Abstract—The local statistical properties of photographic images, when represented in a multiscale basis, have been described using Gaussian scale mixtures. Here, we use this local description as a substrate for constructing a global field of Gaussian scale mixtures (FoGSM). Specifically, we model multiscale subbands as a product of an exponentiated homogeneous Gaussian Markov random field (hGMRF) and a second independent hGMRF. We show that parameter estimation for this model is feasible and that samples drawn from a FoGSM model have marginal and joint statistics similar to those of the subband coefficients of photographic images. We develop an algorithm for removing additive white Gaussian noise based on the FoGSM model and demonstrate denoising performance comparable with state-of-the-art methods.

Index Terms—Image statistics, Markov random field, image denoising.

✦

---

1 INTRODUCTION

MANY successful methods in image processing and computer vision rely on statistical models for images and it is of continuing interest to develop improved models, both in terms of their ability to precisely capture image structures and their practicality for use in applications. A common method of constructing such statistical models is to first identify statistical properties of photographic images and then develop probabilistic models that capture these properties. The first step in this process is to choose a representation (typically, a linear basis) in which the statistical properties are better revealed. Early research in image statistics was based primarily on pixel and Fourier representations. But over the past two decades, numerous studies have demonstrated that linear image decompositions based on multiscale oriented localized basis functions (loosely referred to as "wavelets") are particularly effective in revealing statistical regularities of photographic images. For instance, multiscale subband coefficients of photographic images generally have highly kurtotic non-Gaussian marginal distributions [1], [2], [3] and the *amplitudes* of nearby coefficients are strongly correlated [4], [5], [6].

A variety of parametric models has been proposed to capture these regularities, including the generalized Laplacian [7], [8], [9], [10], the Bessel K [11], the multivariate Student's t-distribution [12], the α-stable family [13], and the Cauchy distribution [14]. All of these non-Gaussian statistical models can be unified under a flexible semiparametric density family known as Gaussian scale mixtures

(GSMs) [15], [16]. By definition, a GSM density is an infinite mixture of zero-mean Gaussian variables with covariances related by multiplicative scaling. GSM can emulate many of the non-Gaussian statistical behaviors observed in local groups of subband coefficients of photographic images. In addition, the underlying Gaussian structure leads to relatively simple parameter learning and inference procedures. For these reasons, local image models based on GSM have been highly successful when applied to image denoising [17], [18].

Despite this success, it has proven difficult to extend the local GSM description to a consistent global probability model. One can partition the coefficient space into nonoverlapping blocks and describe each of these using an independent GSM. But, such a model will ignore important statistical dependencies between coefficients in adjacent blocks. The inhomogeneities that arise from treating coefficients near block boundaries differently from those in the center can, in turn, lead to noticeable artifacts such as blockiness or aliasing in applications such as denoising. This problem may be somewhat ameliorated by using overlapping (e.g., convolutional) blocks [17], [18]. But then, treating these blocks as independent samples is not consistent with any global model. Another option is to retain nonoverlapping coefficient blocks but to capture the dependencies between these blocks by linking the hidden scaling variables in a tree-structured Markov model (e.g., [19], [20]). Although these models are able to capture some global statistical dependencies, they still produce artifacts due to the inhomogeneous treatment of spatially proximal coefficients that are assigned to different branches of the tree.

A natural means of extending the local GSM description to a homogeneous global description is through the use of Markov random fields (MRFs). An MRF is a global model uniquely determined by a set of local statistical descriptions. A number of authors have developed MRF-based image models in the pixel domain (e.g., [21], [22], [23], [24], [25], also see [26] for an overview). In particular, the recently developed field of experts (FoE) model [27] has been used to achieve impressive performance in denoising. However, these MRF-based models usually involve learning and inference

• S. Lyu is with the Computer Science Department, University at Albany, State University of New York, LI-67A, 1400 Washington Avenue, Albany, NY 12222. E-mail: lsw@cs.albany.edu.

• E.P. Simoncelli is with the Howard Hughes Medical Institute, Center for Neural Science, and the Courant Institute of Mathematical Sciences, New York University, 4 Washington Place, Rm. 809, New York, NY 10003-1056. E-mail: eero.simoncelli@nyu.edu.
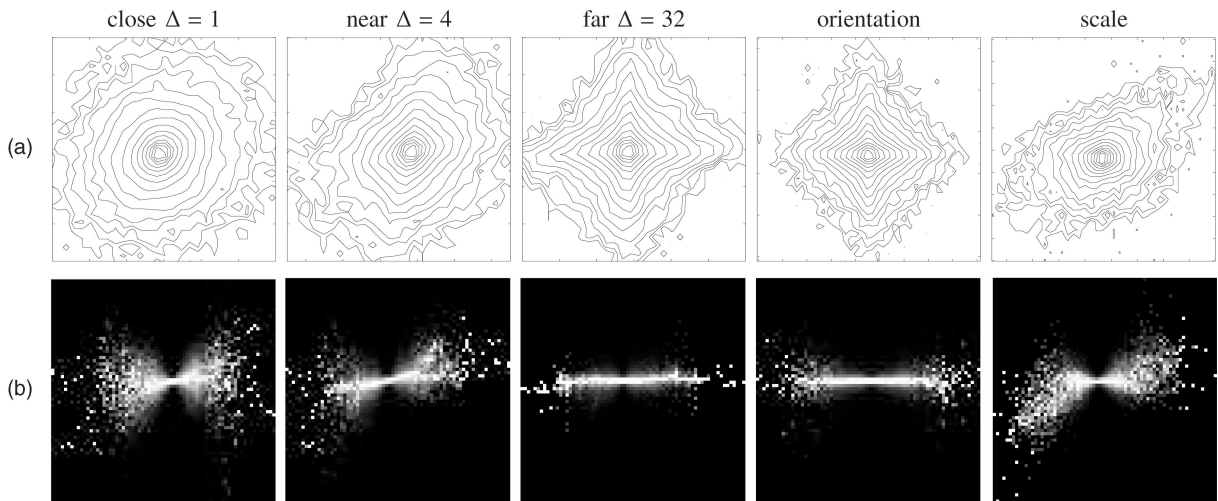
Fig. 1. Histograms of pairs of subband coefficients from the steerable pyramid decomposition of the "boat" image. (a) Contour plots of joint histograms, drawn at equal intervals of log probability. (b) Conditional histograms, computed by independently normalizing each column of the joint histogram. Image intensities are proportional to probability, except that each column of pixels is independently rescaled so that the largest probability value is white.

procedures that are generally computationally costly or unstable.

In this paper, we take a different approach to embedding a local GSM description within a global consistent MRF, by modeling multiscale subbands as *fields of Gaussian scale mixtures* (FoGSMs). Specifically, a FoGSM is formed by an element-wise product of two mutually independent MRFs: a homogeneous Gaussian MRF (hGMRF) and a positive-valued MRF obtained by exponentiating a second hGMRF. The former captures second-order dependencies, while the latter characterizes the variability and dependencies of local variance. Individual coefficients in the FoGSM model marginally follow a GSM distribution, while the global MRF structure generates dependencies beyond local neighborhoods. We develop a parameter estimation procedure, exploiting the computational advantages of the underlying hGMRFs, and demonstrate that samples from FoGSMs share important statistical properties of photographic images. As a sample application, we develop a Bayesian denoising methodology using FoGSM as a prior model for clean images. We show that the resulting denoising method achieves performance comparable to state-of-the-art methods. Preliminary results of this work have been presented in [28].

2 BACKGROUND

2.1 Photographic Image Statistics

Photographic images exhibit distinct statistical regularities that are especially apparent when they are represented using a multiscale oriented decomposition (loosely referred to as a "wavelet" decomposition). To be more specific, the multiscale coefficients of photographic images tend to have highly kurtotic non-Gaussian marginal distributions [1], [2], [3]. More importantly, even when they are second-order decorrelated, there are higher order statistical dependencies between coefficients at nearby locations, orientations, and scales [5], [6], [16]. Fig. 1 shows the empirical joint and conditional histograms for five pairs of subband coefficients of the "boat" image, corresponding to basis functions with spatial separations of $\Delta = \{1, 4, 32\}$ samples, two orthogonal orientations, and two adjacent scales. For spatially

adjacent coefficients (first column), we observe an approximately elliptical joint distribution. This behavior was originally reported for Hilbert-transform pairs of basis functions [30] and later generalized to pairs at different locations, orientations, and scales [5], [6]. The "bow-tie" shaped conditional distribution indicates that the variance of one coefficient depends on the value of the other. This is a highly non-Gaussian behavior, since the conditional variances of a jointly Gaussian density are always constant, independent of the value of the conditioning variable. For coefficients that are distant, the dependency becomes weaker and the corresponding joint and conditional histograms become more separable, as would be expected for two independent random variables. Finally, although the examples shown here were generated using a particular multiscale oriented image representation, these statistical properties are fairly robust to the specific choice of decomposition as long as the basis functions are localized and band-pass.

2.2 Gaussian Scale Mixtures

A GSM vector is defined as the product of a zero-mean Gaussian vector and an independent positive scalar variable. Specifically, a d-dimensional GSM vector $\mathbf{x}$ can be constructed as $\mathbf{x} = \sqrt{z} \cdot \mathbf{u}$, where $\mathbf{u}$ is a d-dimensional zero-mean Gaussian vector and $z \in \mathcal{R}^+$ is independent of $\mathbf{x}$. The density of $\mathbf{x}$ is determined by the covariance matrix, Σ, of the Gaussian vector and the density of z:

$$
\begin{aligned}
p(\mathbf{x}) &= \int_z \mathcal{N}_{\mathbf{x}}(0, z\Sigma) p_z(z) dz \\
&= \int_z \frac{1}{\sqrt{(2\pi z)^d |\Sigma|}} \exp\left(-\frac{\mathbf{x}^T \Sigma^{-1} \mathbf{x}}{2z}\right) p_z(z) dz.
\end{aligned}
\tag{1}
$$

As a family of probability densities, GSM includes many common kurtotic distributions, including all those mentioned in Section 1 [15]. For instance, if z follows an inverse gamma distribution, the resulting GSM density reduces to a multivariate Student's t-distribution [15], [31].

2.3 Homogeneous Gauss-Markov Random Fields

A Markov random field (MRF) is a global joint distribution on a mesh of nodes that is uniquely determined by the local density of each node conditioned on the nodes in a surrounding neighborhood. In particular, the MRF is the maximal entropy density consistent with the local probabilistic constraints [26]. A Gaussian MRF (GMRF) is one in which all of the local conditional (and, hence, joint) densities are Gaussian. In this case, the inverse covariance matrix (also known as the *precision* matrix) of the full set of nodes contains a zero entry for all pairs of nodes that are not within each other's conditioning neighborhoods. The sparse form of the precision matrix means that it usually provides a more convenient parameterization of a GMRF than the full covariance matrix. A homogeneous GMRF (hGMRF) is a GMRF with local density parameters invariant to absolute spatial location. In particular, when the hGMRF is defined over a 2D lattice with circular boundary handling,[1] its precision matrix is block-circulant (see Appendix A for details), determined by a generating kernel Q that captures nonzero dependencies within each neighborhood. Given the relatively small set of parameters, the block-circulant structure, and the resulting close relationship with the discrete Fourier transform (DFT; see Appendix A), hGMRFs are significantly more tractable than general MRFs in terms of parameter estimation, sampling, and inference. Learning and sampling with hGMRFs are described in Appendices B and C, respectively, and a more detailed description of GMRFs and hGMRFs may be found in [32].

3 FIELDS OF GAUSSIAN SCALE MIXTURES

The GSM model has been used successfully to describe the statistics of local clusters of multiscale subband coefficients, which can include spatial neighbors as well as coefficients in adjacent scale and orientation subbands (e.g., [19]). But, as mentioned in Section 1, extending local GSM model to a global model of images without introducing either statistical inconsistencies or inhomogeneities in the global model structure is difficult. Here, we resolve this dilemma by describing each subband as a homogeneous FoGSM.

We define a FoGSM as the element-wise product of two mutually independent MRFs, $\mathbf{u}$ and $\sqrt{\mathbf{z}}$:

$$\mathbf{x} \stackrel{d}{=} \mathbf{u} \otimes \sqrt{\mathbf{z}}, \qquad (2)$$

where the square root is applied to each component of $\mathbf{z}$. Here, $\mathbf{u}$ is a zero-mean GMRF, and $\mathbf{z}$ is a positive-valued MRF of scaling variables. To eliminate the scaling ambiguity between $\mathbf{u}$ and $\mathbf{z}$, we assume that each component of $\mathbf{u}$ has unit variance. The FoGSM model inherits from the GSM model the construction as a product of an independent Gaussian variable and another positive random variable, and as such, all 1D marginal densities of a FoGSM are GSMs. But, unlike the local GSM model, in which a single z variable is multiplied by every component of a multivariate Gaussian variable $\mathbf{u}$, the Gaussian components in FoGSM each have their own z variable. This collection of z variables form a second MRF, which can capture higher order statistical dependencies.

---

1. Circular boundary handling is assumed in the definition of hGMRF. As the dimensionality of the random field increases, the boundary handling is less influential in the computation.

To reduce the number of free parameters in the model, we use *homogeneous* FoGSMs to model each subband in a multiscale decomposition. Specifically, we assume $\mathbf{u}$ to be a zero-mean hGMRF, with circular boundary handling:

$$p(\mathbf{u}) \propto \sqrt{|\mathcal{C}(Q_u)|} \exp\left(\frac{-\mathbf{u}^T \mathcal{C}(Q_u)\mathbf{u}}{2}\right), \qquad (3)$$

where Q_u is the generating kernel and $\mathcal{C}(Q_u)$ is the block-circulant precision matrix formed from that kernel. Furthermore, we assume that $\mathbf{z}$ is derived by applying a point-wise exponential "link" function to a second hGMRF. Equivalently, $\log \mathbf{z}$ (where the $\log$ operator is applied element-wise) is a zero-mean hGMRF with precision matrix $\mathcal{C}(Q_z)$:

$$p_{\mathbf{z}}(\mathbf{z}) \propto \frac{\sqrt{|\mathcal{C}(Q_z)|}}{\prod_i z_i} \exp\left(\frac{-(\log \mathbf{z})^T \mathcal{C}(Q_z)(\log \mathbf{z})}{2}\right). \qquad (4)$$

The interdependencies between components of $\mathbf{z}$ may be explicitly incorporated through the precision matrix $\mathcal{C}(Q_z)$. This lognormal random field is a natural extension of the univariate lognormal density used previously for the scalar multiplier in a local GSM model [33].

The density of $\mathbf{x}$ conditioned on $\mathbf{z}$ may be easily written by substituting the element-wise quotient $\mathbf{x} \oslash \sqrt{\mathbf{z}}$ for the vector $\mathbf{u}$ in (6) and renormalizing:

$$
\begin{aligned}
p(\mathbf{x}|\mathbf{z}) &\propto \sqrt{\frac{|\mathcal{C}(Q_u)|}{\prod_i z_i}} \exp\left(\frac{-(\mathbf{x} \oslash \sqrt{\mathbf{z}})^T \mathcal{C}(Q_u)(\mathbf{x} \oslash \sqrt{\mathbf{z}})}{2}\right) \\
&= \sqrt{\frac{|\mathcal{C}(Q_u)|}{\prod_i z_i}} \exp\left(\frac{-\mathbf{x}^T \mathcal{D}(\sqrt{\mathbf{z}})^{-1} \mathcal{C}(Q_u)\mathcal{D}(\sqrt{\mathbf{z}})^{-1}\mathbf{x}}{2}\right),
\end{aligned}
$$
$$\qquad (5)$$

where $\mathcal{D}(\sqrt{\mathbf{z}})$ in the second line denotes a square diagonal matrix generated from vector $\sqrt{\mathbf{z}}$. The resulting conditional density on $\mathbf{x}$ is a zero-mean *inhomogeneous* GMRF, as its precision matrix $[\mathcal{D}(\sqrt{\mathbf{z}})^{-1}\mathcal{C}(Q_u)\mathcal{D}(\sqrt{\mathbf{z}})^{-1}]$ no longer has a block-circulant structure.

3.1 Learning and Sampling FoGSMs

A FoGSM density on subband coefficients $\mathbf{x}$ is determined by the generating kernels of the two constituent hGMRFs, Q_u and Q_z. When fitting FoGSM to data, it is also desirable to have an estimate of the field $\mathbf{z}$. Thus, we formulate the learning of FoGSM as simultaneous estimation of parameters Q_z and Q_u and variables $\{\mathbf{z}_i\}_{i=1}^N$, from a training set of subbands $\{\mathbf{x}_i\}_{i=1}^N$, as

$$\underset{\{\mathbf{z}_i\}_{i=1}^N, Q_u, Q_z}{\arg\max} \ \log p\left(\{\mathbf{x}_i\}_{i=1}^N, \{\mathbf{z}_i\}_{i=1}^N; Q_u, Q_z\right). \qquad (6)$$

Optimization of this objective function corresponds to a combination of *maximum likelihood* estimation of the model parameters (Q_z, Q_u) and *maximum a posteriori (MAP)* estimation of the hidden variables $\{\mathbf{z}_i\}_{i=1}^N$ from training data $\{\mathbf{x}_i\}_{i=1}^N$.

We optimize (6) using a coordinate ascent scheme, which alternates between maximizing each of $\{\mathbf{z}_i\}_{i=1}^N$, Q_u, and Q_z while holding the remaining two fixed:

1. $\mathbf{z}_i^{(t+1)} = \operatorname{argmax}_{\mathbf{z}} \log p\left(\mathbf{x}_i, \mathbf{z}_i; Q_u^{(t)}, Q_z^{(t)}\right)$ for $i = 1, \ldots, N$,

2. $Q_u^{(t+1)} = \operatorname{argmax}_{Q_u} \log p\left(\left\{\mathbf{x}_i, \mathbf{z}_i^{(t+1)}\right\}_{i=1}^N; Q_u, Q_z^{(t)}\right)$,

3. $Q_z^{(t+1)} = \operatorname{argmax}_{Q_z} \log p\left(\left\{\mathbf{x}_i, \mathbf{z}_i^{(t+1)}\right\}_{i=1}^N; Q_u^{(t+1)}, Q_z\right)$.

$$(7)$$

Running the three steps in (7) iteratively guarantees convergence to a local maximum of the objective function in (6).

The objective function in step 1 in (7) is more conveniently expressed in terms of the element-wise inverse square root of variable $\mathbf{z}$. We define $\mathbf{s} = 1 \oslash \sqrt{\mathbf{z}}$, from which $\mathbf{z}$ can be recovered using $\mathbf{z} = 1 \oslash (\mathbf{s} \otimes \mathbf{s})$. The conditional density of $\mathbf{x}$ given $\mathbf{s}$ may then be written as

$$p(\mathbf{x}|\mathbf{s}) \propto \prod_i s_i \exp\left(\frac{-(\mathbf{x} \otimes \mathbf{s})^T \mathcal{C}(Q_u)(\mathbf{x} \otimes \mathbf{s})}{2}\right)$$
$$= \prod_i s_i \exp\left(\frac{-\mathbf{s}^T [\mathcal{D}(\mathbf{x})\mathcal{C}(Q_u)\mathcal{D}(\mathbf{x})]\mathbf{s}}{2}\right), \tag{8}$$

and the density of $\mathbf{s}$ may be easily obtained by suitable transformation of the density of $\mathbf{z}$:

$$p(\mathbf{s}) \propto \frac{1}{\prod_i s_i} \exp\left(-2(\log \mathbf{s})^T \mathcal{C}(Q_z) (\log \mathbf{s})\right). \tag{9}$$

Using these new definitions, step 1 in (7) may be rewritten as

$$\hat{\mathbf{s}} \overset{\triangle}{=} \operatorname*{argmax}_{\mathbf{s}} \log p(\mathbf{x}, \mathbf{s}; Q_u, Q_z)$$
$$= \operatorname*{argmax}_{\mathbf{s}} \log p(\mathbf{x}|\mathbf{s}; Q_u) + \log p(\mathbf{s}; Q_z)$$
$$= \operatorname*{argmax}_{\mathbf{s}} \begin{cases} \frac{1}{2}\mathbf{s}^T [\mathcal{D}(\mathbf{x})\mathcal{C}(Q_u)\mathcal{D}(\mathbf{x})]\mathbf{s} \\ \quad + 2(\log \mathbf{s})^T \mathcal{C}(Q_z) (\log \mathbf{s}). \end{cases} \tag{10}$$

This objective function may be optimized with conjugate gradient descent [34]. Much of the computation involves multiplying vectors by the precision matrix. Because the precision matrix is block-circulant, these operations are convolutions and may be efficiently implemented using the fast Fourier transform. Empirically, we also found that the conjugate gradient iteration converges quickly: After roughly 300 steps of iteration for a 512×512 pixel image, the successive relative changes in the objective function are less than 10^{-13}.

Steps 2 and 3 in (7) correspond to estimating model parameters Q_u and Q_z given data $\{\mathbf{x}_i, \mathbf{z}_i\}_{i=1}^N$. Specifically, step 2 may be simplified to

$$\hat{Q}_u = \operatorname*{argmax}_{Q_u} \log p(\mathbf{x}, \mathbf{z}; Q_u, Q_z)$$
$$= \operatorname*{argmax}_{Q_u} \log p(\mathbf{x}|\mathbf{z}; Q_u) \tag{11}$$
$$= \operatorname*{argmax}_{Q_u} \log p(\mathbf{x} \oslash \mathbf{z}; Q_u),$$

where the last line corresponds to a *maximum likelihood* estimate of the generating kernel Q_u of a zero-mean hGMRF given N independent samples $\{\mathbf{x}_i \oslash \sqrt{\mathbf{z}_i}\}_{i=1}^N$.

Similarly, step 3 may be simplified as

$$\hat{Q}_z = \operatorname*{argmax}_{Q_z} \log p(\mathbf{x}, \mathbf{z}; Q_u, Q_z)$$
$$= \operatorname*{argmax}_{Q_z} \log p(\mathbf{z}; Q_z) \tag{12}$$
$$= \operatorname*{argmax}_{Q_z} \log p(\log \mathbf{z}; Q_z),$$

which is the *maximum likelihood* estimate of parameter Q_z in an hGMRF on $\log \mathbf{z}$ given independent samples $\{\log \mathbf{z}_i\}_{i=1}^N$. Again, both steps allow efficient computation based on the properties of block-circulant matrices. The details of parameter estimation for hGMRFs are provided in Appendix B.

Sampling from FoGSM is simple and efficient. By definition, a sample of FoGSM is formed by element-wise multiplication of two independent samples of $\mathbf{u}$ and $\sqrt{\mathbf{z}}$. The former is obtained by sampling from hGMRF $\mathbf{u}$ and the latter is obtained by element-wise exponentiation of the square root of a sample of hGMRF $\log \mathbf{z}$. Sampling from each 2D hGMRF is implemented by linearly transforming a sample of white Gaussian noise, which is again efficient due to the computational advantages of block-circulant matrices. We provide basic descriptions of these operations in Appendix C and more information can be found in [32].

4 MODELING MULTISCALE IMAGE SUBBANDS WITH FoGSMs

In this section, we investigate empirically how well an FoGSM is able to account for the statistical properties of subband coefficients of photographic images. We fit an independent FoGSM model to each subband of a photographic image and examine the properties of the $\mathbf{u}$ and $\log \mathbf{z}$ fields, as well as samples from the learned FoGSM model.

4.1 Experimental Setup

Our data sets are multiscale subbands of a given orientation and scale from five standard test images of size 512×512 pixels (known as "Lena," "Barbara," "boats," "hill," and "baboon"). For image representation, we employed an overcomplete tight frame representation known as the steerable pyramid [29]. The basis functions of this linear decomposition are spatially localized, oriented, and span roughly one octave in bandwidth. They are polar separable in the Fourier domain and are related by translation, dilation, and rotation. We fit the FoGSM model to subbands corresponding to the first scale and third orientation in an eight-orientation decomposition (the peak orientation angle of this band is at $\pi/4$ radians, relative to the horizontal axis).

The neighborhood sizes of the two component hGMRFs of the FoGSM model (corresponding to variables $\mathbf{u}$ and $\log \mathbf{z}$) were chosen by maximizing the cross-validated likelihood. We cut each subband into equal-sized rectangular halves, fitted the FoGSM model of a given neighborhood size to one half of the subband, and then computed the likelihood on the data from the other half of the subband using (6). The best performance was observed for 5×5 neighborhoods, for both hGMRFs. Once the neighborhood size was determined, the generating kernels were optimized using the algorithm described in the previous section. The vector $\mathbf{z}$, which represents the local signal variance, was initialized by
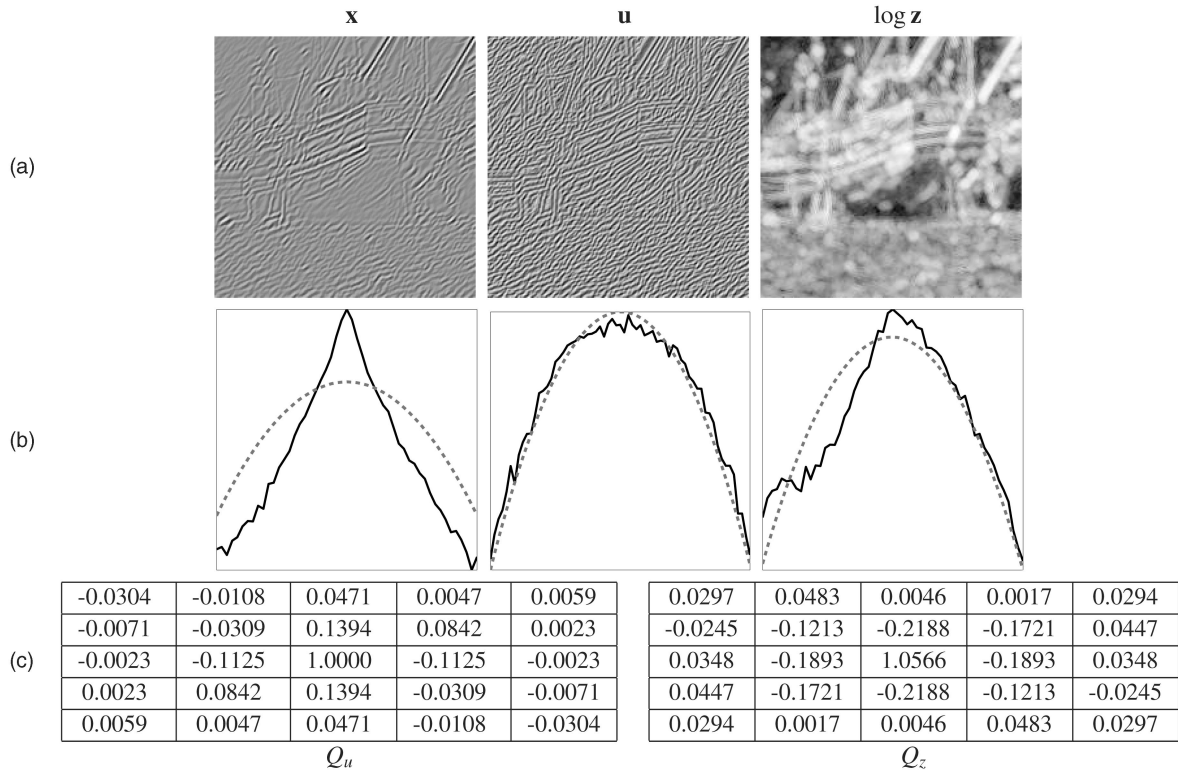
Fig. 2. (a) Decomposition of a subband (left) from image "boat" into an hGMRF $\mathbf{u}$ (middle) and the corresponding multiplier field $\log \mathbf{z}$ (right). Each image is rescaled individually to fill the full range of grayscale intensities. (b) Log marginal histograms of $\mathbf{x}$, the estimated $\mathbf{u}$, and the estimated $\log \mathbf{z}$. Dotted lines correspond to Gaussian density of the same mean and variance. (c) 5×5 central nonzero regions for the hGMRF generating kernels of the estimated $\mathbf{u}$ and $\log \mathbf{z}$ fields.

computing the local variances estimated within each overlapping 5×5 spatial window.

4.2 Decomposition and Parameters

Fig. 2a shows the results of decomposing a subband from the "boats" image according to the fitted FoGSM model. Specifically, the subband (left panel) is decomposed into the product of the $\mathbf{u}$ field (middle panel) and the $\sqrt{\mathbf{z}}$ field (right panel, in logarithm) using the training algorithm described in Section 3.1. Visually, we can see that the changing spatial variances are captured by the estimated $\log \mathbf{z}$ and residual homogeneous structures are captured by the estimated $\mathbf{u}$. Fig. 2b shows the marginal histograms of each field in the log domain, plotted against a Gaussian density of the same variance. Note that the marginal distribution of $\mathbf{u}$ is well approximated by a Gaussian, as assumed in the FoGSM model. The marginal distribution of $\log \mathbf{z}$, while unimodal, exhibits noticeable deviations from Gaussianity. In particular, it is clearly asymmetric, and this property seems to be consistent across a variety of different subbands and images.

Fig. 2c shows the estimated 5×5 generating kernels Q_u and Q_z. The former reflects the orientation anisotropy of hGMRF $\mathbf{u}$, which is matched to the orientation tuning of the subband. On the other hand, Q_z shows only weak orientation preference. We could interpret this to indicate that the MRF for $\log \mathbf{z}$ is close to isotropic. However, visual inspection of the $\log \mathbf{z}$ field suggests that the MRF frequently exhibits strongly oriented content, but that this content is inhomogeneous (i.e., the orientation is different in different image regions) and thus cannot be captured by an

hGMRF. Furthermore, one can see that the estimated $\mathbf{u}$ and $\log \mathbf{z}$ are not independent, as assumed by the FoGSM model, but have aligned structures (typically arising from image contours).

4.3 Statistics of FoGSM Samples

The statistical dependencies captured by the FoGSM model can be further illustrated by examining marginal and joint statistics of samples from the fitted model. Note that this is achieved by fitting the global FoGSM statistical model to a subband and then drawing samples from this model, not by explicitly adjusting parameters to fit the marginal or joint histograms (as was done in [16]).

We begin by comparing the marginal distributions of the samples and the original subband. Fig. 3 shows empirical histograms in the log domain of a particular subband from four different photographic images (dashed line) and those of the synthesized samples of FoGSM models learned from each corresponding subband (solid line). For comparison, a Gaussian with matching standard deviation is also displayed (thin dashed line). Note that the synthesized samples have conspicuous non-Gaussian marginal characteristics, exemplified by the high peak and heavy tails, similar to the image subbands. On the other hand, the synthesized coefficients are typically less kurtotic than the real subbands. The shape of these marginal densities is dictated by the $\mathbf{z}$ field, which is an hGMRF transformed with a point-wise exponential link function. An alternative choice of link function could be used to create distributions closer to the observed multiscale subbands.

In addition to the marginal statistics, the FoGSM model also has joint behaviors that are similar to those observed in
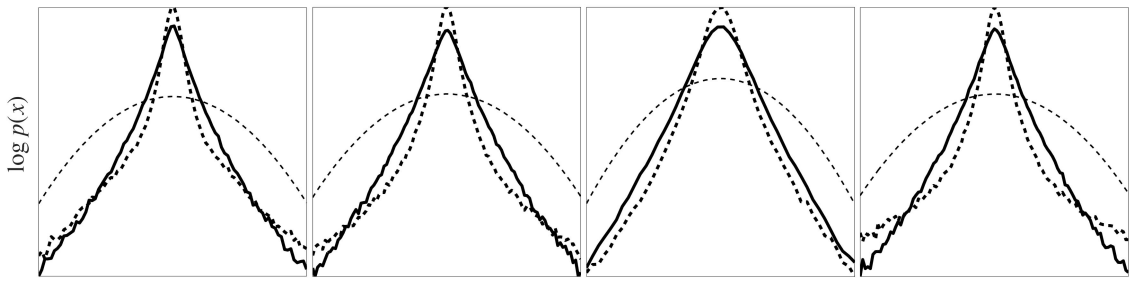
Fig. 3. Marginal log distributions of coefficients from a multiscale decomposition of four photographic images (dashed line), synthesized FoGSM samples from the same subband (solid line), and a Gaussian with the same standard deviation (thin dashed line).
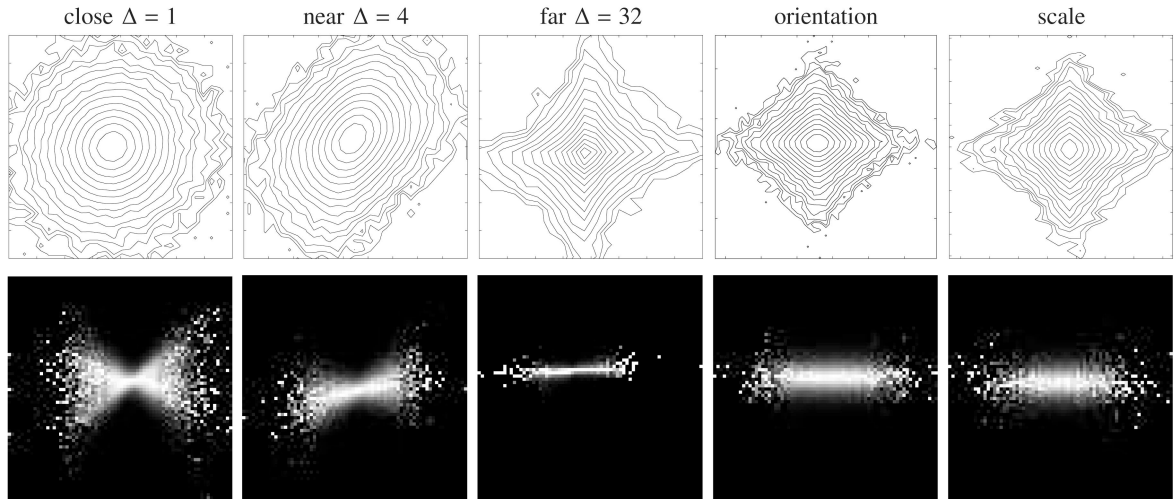

Fig. 4. Joint distribution of pairs of subband coefficients obtained from samples drawn from the best-fitting FoGSM model. See the caption of Fig. 1 for explanation.

multiscale coefficients of photographic images. Fig. 4 shows the joint and conditional histograms of synthesized samples from the FoGSM model estimated from the same subband used to generate the histograms in Fig. 1. Note that histograms of the synthesized samples have a dependence on spatial proximity similar to those of the image data shown in Fig. 1. This behavior arises directly from the structure of the FoGSM model. The random field $\mathbf{z}$ is smooth and, thus, nearby components have nearly identical marginal variance, resulting in an elliptically contoured joint density and strong dependency between coefficients. This dependency is propagated from neighborhood to neighborhood in the FoGSM model but weakens with distance. On the other hand, note that the dependencies between coefficients representing different orientations or scales are not properly modeled because we have used an independent FoGSM to model each subband. This is evident when comparing the fourth and the fifth columns of Figs. 4 and 1.

Finally, Fig. 5 shows samples of $\mathbf{u}$, $\log \mathbf{z}$, and $\mathbf{x}$ drawn according to an FoGSM model whose parameters were fit to the subband shown in Fig. 2. The $\mathbf{u}$ field resembles that of the subband, but the $\log \mathbf{z}$ field is seen to lack the extended structures seen in the data. Thus, the FoGSM model fails to fully capture the inhomogeneous long-range interactions that arise in images around contours or extended features.

5 APPLICATION TO IMAGE DENOISING

As a probability model for photographic images, FoGSM may be used as a prior for Bayesian estimation of an image given an observation corrupted by additive white Gaussian noise of known variance. In addition to its practical relevance, image denoising is a simple yet powerful test for the effectiveness of an image model, providing a clear quantitative test of how well the model can differentiate photographic image content from noise.

5.1 Algorithm

We follow a conventional methodology, decomposing the noise-corrupted image into multiscale subbands, computing an estimate of the coefficients of each subband using the FoGSM model as a prior, and then generating the denoised image by applying the inverse multiscale transform to the denoised subbands. Since the multiscale decomposition is linear, we may write $\mathbf{y} = \mathbf{x} + \mathbf{w}$ for a multiscale subband of the noisy image, where $\mathbf{x}$ is the clean subband and $\mathbf{w}$ is the noise that is added to the subband. Note that, in an overcomplete representation such as steerable pyramid, white Gaussian noise in the image domain is transformed into correlated Gaussian noise, whose covariance can be computed from the basis functions of the transform.

A standard approach to denoising is to formulate it as a Bayesian inference problem, selecting an estimate based on the posterior density $p(\mathbf{x}|\mathbf{y})$, which is proportional to the product of the likelihood function $p(\mathbf{y}|\mathbf{x})$ and the image prior $p(\mathbf{x})$. Two solutions are common. The MAP estimate is the
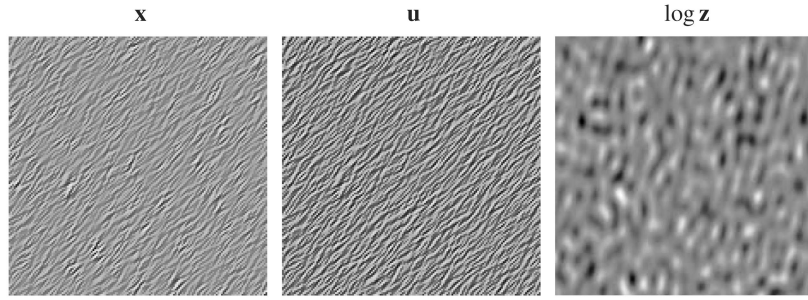
Fig. 5. Samples from the MRFs with parameters shown in Fig. 2.

mode of the posterior density, $\hat{\mathbf{x}}_{MAP} = \operatorname{argmax}_\mathbf{x} \log p(\mathbf{x}|\mathbf{y})$, whereas the *Bayesian least square* (BLS) estimate, which minimizes the expected square error between the restored image and the original image, is the mean of the posterior density $\hat{\mathbf{x}}_{BLS} = E(\mathbf{x}|\mathbf{y})$. Both of these solutions involve computationally expensive (high-dimensional) integration when used with FoGSM model. Specifically, MAP requires a high-dimensional integral over $\mathbf{z}$, while BLS requires high-dimensional integrals over both $\mathbf{x}$ and $\mathbf{z}$.

Although it is possible to obtain approximations to these solutions using Markov chain Monte Carlo [26] or variational approximations [35], we instead develop a deterministic algorithm that takes advantage of the hGMRF structure of the FoGSM model. Specifically, we compute

$$(\hat{\mathbf{x}}, \hat{\mathbf{z}}) = \operatorname*{argmax}_{\mathbf{x}, \mathbf{z}} \log p(\mathbf{x}, \mathbf{z}|\mathbf{y}) \qquad (13)$$

and then take $\hat{\mathbf{x}}$ as the denoised subband. This strategy, known as a "partial optimal solution" [36], greatly reduces the computational complexity of the problem. The solution to the optimization problem in (13) is found by coordinate ascent. Starting with initial values for $\mathbf{x}$ and $\mathbf{z}$, the algorithm proceeds by alternating between the following steps.

5.1.1 Optimization of $\mathbf{x}$

Given the current estimate of $\mathbf{z}$, the optimization of $\mathbf{x}$ in (13) can be expanded using Bayes' rule:

$$\operatorname*{argmax}_{\mathbf{x}} \log p(\mathbf{x}, \mathbf{z}|\mathbf{y})$$
$$= \operatorname*{argmax}_{\mathbf{x}} \{\log p(\mathbf{y}|\mathbf{x}, \mathbf{z}) + \log p(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{z}) - \log p(\mathbf{y})\},$$
$$= \operatorname*{argmax}_{\mathbf{x}} \{\log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x}|\mathbf{z})\},$$

where the first term is simplified because $\mathbf{y}$ and $\mathbf{z}$ are independent when conditioned on $\mathbf{x}$ and the last two terms are dropped because they do not depend on $\mathbf{x}$. Given the Gaussian structure of the first two terms, the maximum is linear in $\mathbf{y}$ (equivalent to a Wiener filter). Specifically, we must minimize a quadratic expression:

$$\operatorname*{argmin}_{\mathbf{x}} \left\{ \begin{array}{l} (\mathbf{y} - \mathbf{x})^T \mathcal{C}(C_w)^{-1}(\mathbf{y} - \mathbf{x}) \\ + \mathbf{x}^T \mathcal{D}(\sqrt{\mathbf{z}})^{-1} \mathcal{C}(Q_u) \mathcal{D}(\sqrt{\mathbf{z}})^{-1}\mathbf{x}, \end{array} \right.$$

where the noise covariance $\mathcal{C}(C_w)$ is a block-circulant matrix determined by a generating kernel C_w that represents the convolution by which the subband is obtained from the image pixels. Note that, although $\mathcal{C}(C_w)$ may be sparse (zero beyond the support of the filters), the inverse of $\mathcal{C}(C_w)$ can still be

dense. It is therefore computationally advantageous to work with $\mathcal{C}(C_w)$ rather than its inverse. For this reason, we introduce $\mathbf{t} = \mathcal{C}(C_w)^{-1}\mathbf{x}$ and find the optimal $\mathbf{t}$ by solving

$$\operatorname*{argmin}_{\mathbf{t}} \left\{ \begin{array}{l} (\mathbf{y} - \mathcal{C}(C_w)\mathbf{t})^T \mathcal{C}(C_w)^{-1}(\mathbf{y} - \mathcal{C}(C_w)\mathbf{t}) \\ + \mathbf{t}^T \mathcal{C}(C_w)^T \mathcal{D}(\sqrt{\mathbf{z}})^{-1} \mathcal{C}(Q_u) \ \mathcal{D}(\sqrt{\mathbf{z}})^{-1} \mathcal{C}(C_w)\mathbf{t} \end{array} \right.$$

or expanding the first line, and dropping the term independent of $\mathbf{t}$:

$$\operatorname*{argmin}_{\mathbf{t}} \left\{ \begin{array}{l} \mathbf{t}^T \mathcal{C}(C_w)\mathbf{t} - 2\mathbf{y}^T\mathbf{t} \\ + \mathbf{t}^T \mathcal{C}(C_w)^T \mathcal{D}(\sqrt{\mathbf{z}})^{-1} \mathcal{C}(Q_u) \ \mathcal{D}(\sqrt{\mathbf{z}})^{-1} \mathcal{C}(C_w)\mathbf{t}. \end{array} \right.$$

Note that this objective function is quadratic in $\mathbf{t}$, guaranteeing a global optimal solution. We compute the optimum using conjugate gradient descent and then recover the optimal $\mathbf{x}$ through the relationship $\hat{\mathbf{x}} = \mathcal{C}(C_w)\hat{\mathbf{t}}$.

5.1.2 Optimization of $\mathbf{z}$

Given the current estimate of $\mathbf{x}$, the optimization of $\mathbf{z}$ in (13) can be written as

$$\operatorname*{argmax}_{\mathbf{z}} \log p(\mathbf{x}, \mathbf{z}|\mathbf{y}) =$$
$$\operatorname*{argmax}_{\mathbf{z}} \{\log p(\mathbf{y}|\mathbf{x}, \mathbf{z}) + \log p(\mathbf{x}, \mathbf{z}) - \log p(\mathbf{y})\}.$$

The last term may be dropped because it is independent of $\mathbf{z}$ and the first term is dropped since $\mathbf{y}$ is independent of $\mathbf{z}$ when conditioned on $\mathbf{x}$. Thus, the problem is reduced to $\operatorname{argmax}_\mathbf{z} \log p(\mathbf{x}, \mathbf{z})$, which may be computed as in step 1 of the learning procedure of Section 3.1.

5.1.3 Acceleration

The alternating optimization of $\mathbf{x}$ and $\mathbf{z}$ is guaranteed to converge to a local optimum of the objective function in (13), but the convergence speed can be very slow. To accelerate convergence, we include a heuristic "inertial" step after every two steps of the optimization loop. Specifically, the algorithm takes a step in the direction established by the optimal values of the previous two iterations, with the step size optimized according to

$$\left\{ \hat{\mathbf{x}}^{(k+1)}, \hat{\mathbf{z}}^{(k+1)} \right\} = \left\{ \mathbf{x_e}^{(k)}(\alpha_{\mathrm{opt}}), \mathbf{z_e}^{(k)}(\alpha_{\mathrm{opt}}) \right\} \text{ where}$$
$$\mathbf{x_e}^{(k)}(\alpha) = \hat{\mathbf{x}}^{(k)} + \alpha \left(\hat{\mathbf{x}}^{(k)} - \hat{\mathbf{x}}^{(k-1)} \right),$$
$$\mathbf{z_e}^{(k)}(\alpha = \hat{\mathbf{z}}^{(k)} + \alpha \left(\hat{\mathbf{z}}^{(k)} - \hat{\mathbf{z}}^{(k-1)} \right),$$
$$\alpha_{\mathrm{opt}} = \operatorname*{argmax}_{\alpha \in (0, \alpha_{max})} \log p(\mathbf{x_e}(\alpha), \mathbf{z_e}(\alpha)|\mathbf{y}).$$

Intuitively, such a jump ensures that the optimization does not oscillate back and forth within a narrow valley of the objective function. In practice, as shown in the following experiments, it achieves a substantial reduction in the overall running time of the algorithm.

5.1.4 Parameter Estimation (Optional)

The denoising algorithm described thus far assumes that the model parameters Q_u and Q_z are known. These model parameters can be learned as a *generic* statistical model for subband coefficients from a large set of noise-free photographic images using the algorithm provided in Section 3.1. The advantage of a generic image model approach is that the training can be performed offline, which may greatly reduce the overall running time. Alternatively, these parameters can be adaptively learned by including a parameter estimation step in the loop of the denoising algorithm

$$\left(\hat{\mathbf{Q}}_{\mathbf{u}}^{(k)}, \hat{\mathbf{Q}}_{\mathbf{z}}^{(k)}\right) = \underset{Q_u, Q_z}{\operatorname{argmax}} \log p\left(\hat{\mathbf{x}}^{(k)}, \hat{\mathbf{z}}^{(k)} | \mathbf{y}; \mathbf{Q}_{\mathbf{u}}, \mathbf{Q}_{\mathbf{z}}\right) \qquad (14)$$

as in Section 3.1. The FoGSM model parameters (Q_u, Q_z) are estimated from hGMRFs $[\mathbf{x} \oslash \sqrt{\hat{\mathbf{z}}}^{(k)}]$ and $[\log \mathbf{z}^{(k)}]$, as in steps 2 and 3 in (7). Adaptively learning the parameters allows the model to better account for the local structure of the image in question, thus potentially leading to better performance. We compare the relative performance of these two training schemes in the following experiments.

5.2 Experimental Setup

We evaluated the FoGSM denoising method on a set of standard grayscale test images [17]. All images are of size 256×256 or 512×512 pixels and in 8-bit TIFF format. Noisy images were generated by adding simulated white Gaussian noise. We evaluate the denoising performance by visual inspection, as well as the conventional objective performance known as peak-signal-to-noise ratio (PSNR), defined as $20 \log_{10}(255/\sigma_e)$, where σ_e is the standard deviation (computed by averaging over spatial position) of the difference between the restored image and the original image.

Each noise-corrupted image was first decomposed into a steerable pyramid with multiple scales (five levels for a 512×512 image and four levels for a 256×256 image) and eight orientations. These values were chosen empirically to achieve a reasonable compromise between denoising performance and computational load. The resulting representation is approximately 11 times overcomplete, relative to the original image size. The Markov neighborhoods for hGMRFs $\mathbf{u}$ and $\log \mathbf{z}$ were both chosen to cover 5×5 blocks of coefficients since this was found to be optimal for representation of clean images. We verified that this specific choice was also roughly optimal for the best denoising performances across different images and noise levels. The model parameters were obtained by training the FoGSM model adaptively for each subband as described in Section 5.1, with initial parameter values chosen to represent a smooth and isotropic GMRF. The initial values of $\mathbf{x}$ and $\mathbf{z}$ are computed from subband denoised with the local GSM model [17].

5.3 Results

Denoising results for six test images, at seven different noise levels, are reported in Table 1. The standard deviations of PSNR values for each image and noise level, computed by repeating each denoising experiment 10 times with different samples of noise, are consistently lower than 0.1 dB. In addition to the results for our FoGSM algorithm, we also provide denoising results of the BLS-GSM method [17]. This algorithm computes the Bayes least squares estimate (i.e., conditional mean) of individual coefficients based on a local GSM model. The comparison with FoGSM allows us to assess the gain in performance that is obtained by building a global model. We employed the implementation described in [17], which assumes a neighborhood consisting of 3×3 spatial neighbors plus a "parent" coefficient in the next coarsest scale.[2] Finally, we provide results of the current state-of-the-art denoising method, BM3D [37]. The PSNR values for these methods were directly taken from their corresponding publications. Note that the FoGSM algorithm achieves consistent improvements in PSNR over the local GSM-based algorithm (average improvement is 0.52 dB), clearly demonstrating the advantage of a globally consistent statistical model. On the other hand, the performance of the FoGSM method is comparable (sometimes better, sometimes worse) to that of BM3D, which is not based on any explicit statistical model. In general, BM3D relies on the image containing repeating patterns (specifically, many blocks of pixels that are similar). Thus, it performs best on images with large regions of the same texture (e.g., "Barbara") or long contours of similar orientation (e.g., "House") and performs less well on images with diverse content (e.g., "boats," or "Lena").

In Fig. 6, we plot the PSNR performance of four recent denoising methods relative to that of FoGSM. The performance of FoGSM is consistently better than that of kSVD [39], BLS-GSM [17], and FoE [27], and (on average) comparable to those of BM3D [37].

Next, we examine and compare the denoising results visually. Figs. 7 and 9 show the results of denoising the "Barbara" image and the "boats" image with noise level $\sigma = 50$, corresponding to a PSNR of 14.15 dB. We have chosen a relatively high level of noise, in order to provide a clear visualization of the capabilities and limitations of the model. To better examine the details of the denoising results, we show in Figs. 8 and 10 enlargements of selected regions of each of the corresponding images in Figs. 7 and 9, respectively. For these examples, the FoGSM denoising achieves substantial improvements (+0.95 and +0.68 dB) and is seen to exhibit higher contrast and continuation of oriented features. However, FoGSM also introduces some noticeable artifacts in low contrast regions, which are likely due to failures of the FoGSM model to capture all statistical properties of photographic image multiscale subband coefficients. For example, coefficient amplitudes are known to be correlated across scale (see Fig. 4, right panel). If represented properly, this correlation should allow the denoising algorithm to recognize isolated large coefficients as noise, since (unlike photographic images) they will not have corresponding large-amplitude coefficients in adjacent bands. But, the current model treats each subband

---

2. A MAP-based denoising algorithm based on the local GSM model was developed in [38]. However, its performance was significantly worse than the BLS method published in [17].

TABLE 1
Comparison of FoGSM with BLS-GSM [17] and BM3D [37]

σ/PSNR	Barbara 512×512			boats 512×512			hill 512×512		
	FoGSM	GSM	BM3D	FoGSM	GSM	BM3D	FoGSM	GSM	BM3D
5/34.15	**38.65**	37.79	38.31	**37.39**	36.97	37.28	**37.16**	36.91	**37.14**
10/28.13	**35.01**	34.03	**34.98**	**34.12**	33.58	33.92	**33.78**	33.38	33.62
15/24.61	32.85	31.86	**33.11**	**32.31**	31.70	32.14	**31.99**	31.51	31.86
25/20.17	30.10	29.13	**30.72**	**30.03**	29.37	29.91	**29.91**	29.37	29.85
50/14.15	26.40	25.48	**27.17**	**27.01**	26.38	26.64	**27.38**	26.82	27.08
75/10.63	24.29	23.65	**25.10**	**25.33**	24.79	24.96	**25.93**	25.46	25.58
100/8.13	23.01	22.61	**23.49**	**24.20**	23.75	23.74	**24.88**	24.53	24.45

σ/PSNR	house 256×256			Lena 512×512			peppers 256×256		
	FoGSM	GSM	BM3D	FoGSM	GSM	BM3D	FoGSM	GSM	BM3D
5/34.15	38.98	38.65	**39.83**	**38.66**	38.49	**38.72**	37.91	37.30	**38.12**
10/28.13	35.63	35.35	**36.71**	**35.94**	35.61	**35.93**	34.38	33.73	**34.68**
15/24.61	33.89	33.64	**34.94**	**34.28**	33.90	**34.27**	32.34	31.70	**32.70**
25/20.17	31.64	31.40	**32.86**	**32.11**	31.69	**32.08**	29.78	29.18	**30.06**
50/14.15	28.51	28.26	**29.37**	**29.12**	28.61	28.86	**26.43**	25.93	**26.41**
75/10.63	26.69	26.41	**27.20**	**27.37**	26.84	27.02	**24.53**	24.11	**24.48**
100/8.13	25.33	25.11	**25.50**	**26.12**	25.64	25.57	**23.17**	22.80	22.91

Performance values are expressed as PSNR, $20 \log_{10}(255/\sigma_e)$, where σ_e is the standard deviation of the difference between the denoised image and the original image. Numbers in boldface indicate the best performance among the three methods for each image and noise level. Cases in which the two best methods differ by less than 0.1 dB are considered a tie.
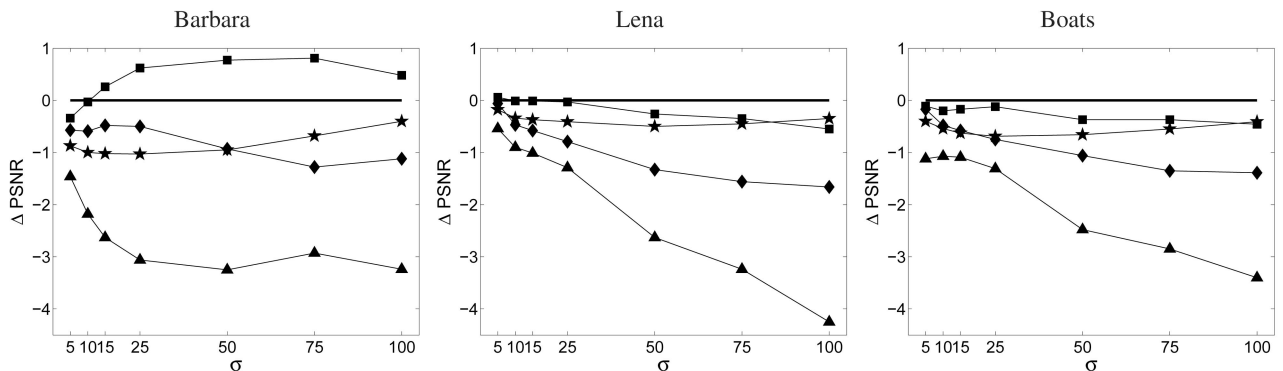

Fig. 6. Performance comparison of denoising methods for three different images. Plotted are differences in PSNR for different input noise levels (σ) between FoGSM and four other methods (square, BM3D [37], star, BLS-GSM [17], diamond, kSVD [39], and triangle, FoE [27]). The PSNR values for these methods were taken from their corresponding publications.

independently, thus allowing these isolated coefficients to remain as unsuppressed artifacts. In addition, these artifacts may also be aggravated by the use of a MAP-like estimator. A local MAP-GSM estimator produces similar unsuppressed coefficients, when compared to the smoother behavior of the local BLS-GSM estimate.

The denoising performance obtained with FoGSM is attained with a substantial computational cost. As a rough indication, our unoptimized MATLAB code, running on an Intel workstation with 2.6 GHz dual Opteron 64-bit processor and 16 Gbyte RAM memory, takes, on average, 97.3 minutes (results averaging over nine trials, with a range of [71.8, 124.4] minutes) to denoise a 512×512 image at noise level $\sigma = 50$, and takes, on average, 35.3 minutes (result averaging over four images, with a range of [28.4, 47.9] minutes) to denoise a 256×256 image at the same noise level. It is likely that these values could be improved by incorporating additional acceleration heuristics.

5.4 Algorithm Variations

In order to understand the contribution of various aspects of the FoGSM-based denoising method, we examined their relative effect on the denoising performance. Table 2 shows the changes in PSNR and running time when various features of the method are modified. All results are for noise level $\sigma = 50$ and averaged over three 512×512 images (they are "boats," "Lena," "Barbara"). The first two columns correspond to modifications of the front-end representation. The first (ortho wvlt) corresponds to using an approximately orthogonal wavelet decomposition based on quadrature mirror filters [40]. The separable QMF pyramid splits the image frequency domain into horizontal, vertical, and mixed diagonal subbands. Using this representation results in a substantial reduction in performance, which we believe is partly due to the mixed orientations in the diagonal band and partly due to the lack of overcompleteness that generally improves denoising [41], [42].

The second column (4 orns) shows the result of using steerable pyramid decomposition with only four orientations (instead of eight). Decreasing the number of orientations

Fig. 7. Denoising results using local GSM [17] and FoGSM.

original image

noisy image ($\sigma = 50$) (PSNR = 14.15dB)

local GSM [17] (PSNR = 25.45dB)

FoGSM (PSNR = 26.40dB)

original image

noisy image ($\sigma = 50$)
(PSNR = 14.15dB)

BLS-GSM
(PSNR = 25.45dB)
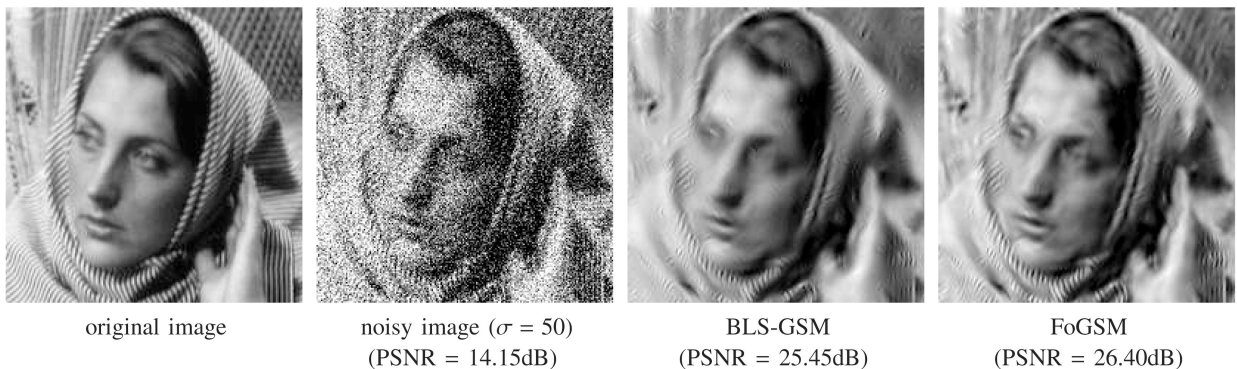
FoGSM
(PSNR = 26.40dB)

Fig. 8. Enlargement of a selected region of the images in Fig. 7.

leads to a significant drop in performance, accompanied by a substantial reduction in running time. On the other hand, increasing the number of orientations (not shown) leads to small improvements in PSNR, at the expense of considerable computation cost.

In the next three columns, we examine the effects in the FoGSM model structure. We first compared the effect of different choices of MRF neighborhood size (3×3 and 7×7). As shown, changing neighborhood size has relatively little effect on the overall running time, but the PSNR values were lower for both neighborhood sizes, justifying our choice of the 5×5 neighborhood (at least for this image and noise level). In the fifth column (*gen param*), we compared the result of the offline training of the (Q_u, Q_z) parameters of the FoGSM

| original image | noisy image ($\sigma = 50$) (PSNR = 14.15dB) |
| local GSM [17] (PSNR = 26.34dB) | FoGSM (PSNR = 27.02dB) |

Fig. 9. Denoising results using local GSM [17] and FoGSM.

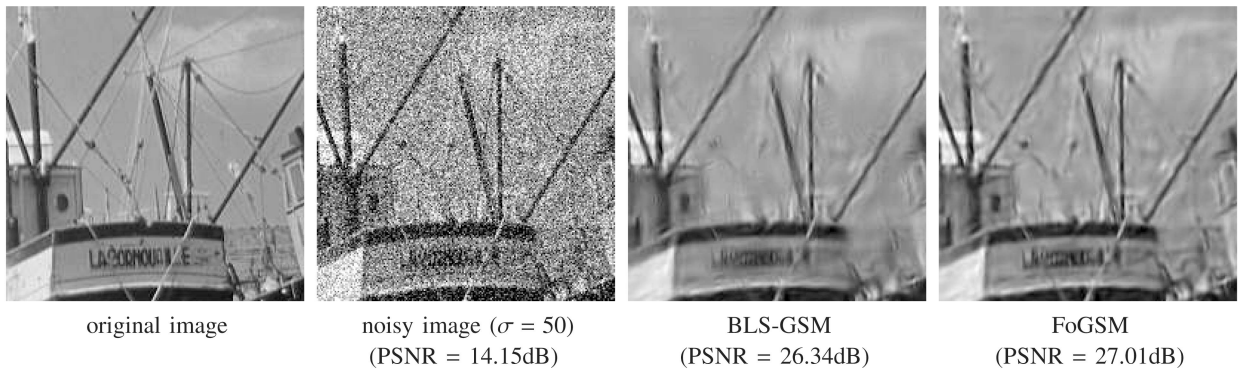| original image | noisy image ($\sigma = 50$) (PSNR = 14.15dB) | BLS-GSM (PSNR = 26.34dB) | FoGSM (PSNR = 27.01dB) |

Fig. 10. Enlargement of a selected region of the images in Fig. 9.

model on a set of noise-free images (not including the three test images). This leads to a significant reduction in the computational cost of denoising an image since the parameter learning step no longer needs to be included in the denoising process. However, this is accompanied by a significant loss in denoising performance (an average PSNR reduction of approximately 0.4 dB), since the generically learned model parameters are less adapted to the idiosyncrasies of the specific image/subband being denoised. The last column (*no accel*) shows that the accelerating heuristics introduced in the previous section significantly improve the running time of the denoising procedure, while having a negligible effect on PSNR.

6 RELATED MODELS

The local GSM model that underlies the FoGSM is closely related to other local hidden variable models for images [16], [8], [43], [31], [44]. However, the use of MRFs in the FoGSM allow it to extend to images of arbitrary size in a statistically consistent way, while the local scale mixture models are essentially confined to describing small image patches. The underlying MRF structure of the hidden variables in the FoGSM model also differentiates it from mixture models with tree-structured hidden variables [19], [20]. These models have the advantage of explicitly

TABLE 2
Effects of Different Algorithm Variations

	ortho wvlt	4 orns	3×3 nb	7×7 nb	gen param	no accl
ΔPSNR (dB)	-0.47	-0.32	-0.07	-0.17	-0.39	0.01
Δt/t (%)	-62.8	-52.6	-1.2	2.5	-41.7	163.8

ΔPSNR *specifies the changes in PSNR resulting from a change in the corresponding attribute.* Δt/t *specifies the percentage of change in running time relative to the running time of the standard algorithm described in Section 5.3. All values are averages over three 512×512 images ("boats," "Lena," "Barbara") for the noise level $\sigma = 50$.*

capturing cross-band dependencies, but they suffer from spatial inhomogeneities introduced by the tree partitioning.

As a global MRF-based image model, the basic architecture of FoGSM differs from existing non-Gausssian MRF image models [21], [23], [24], [27] in that it is not defined by specification of clique potentials but through nonlinear composition of two hGMRFs. On the other hand, FoGSM has some resemblance to the compound GMRFs model for images [22], which is formed by modulating an hGMRF with a binary line process that indicates the existence of an edge between two spatial locations [21]. A modified version, proposed in [36], treats the hidden variables as independent. This simplifies computation but may lead to a loss in performance in applications.

7   DISCUSSION

We have introduced FoGSMs as a flexible and efficient tool for modeling the statistics of multiscale subband coefficients of photographic images. We developed a feasible parameter estimation method and showed that samples synthesized from the fitted FoGSM model are able to capture structures in the marginal and joint subband statistics of photographic images. We have applied FoGSM to image denoising and demonstrated performance comparable to current state-of-the-art denoising methods.

We envision, and are currently working on, a number of improvements. First, the model should benefit from the introduction of more general Markov neighborhoods, including coefficients from subbands at other scales and orientations [6], [17], since the current model is clearly not capturing these dependencies. A natural means of achieving this is to allow different subbands to share the same hidden scaling field, although this may substantially complicate the learning and inference algorithms. A possible remedy is to capture these cross-scale dependencies with a coarse-to-fine conditional model. Second, the logarithmic link function used to derive the multiplier field from an hGMRF was chosen somewhat arbitrarily, and we believe that substitution of another nonlinear transformation (e.g., a power law, as in [19]) could lead to a more accurate description of marginal and coefficient statistics. Finally, there exist residual inhomogeneous structures in both the $\mathbf{u}$ and $\log \mathbf{z}$ fields (see Fig. 2) that may be captured by explicitly incorporating local orientation [45] or phase [46] into the model. Finding tractable models and algorithms for handling such angular variables is challenging, but we believe their inclusion will result in substantial improvements in modeling and in denoising performance.

APPENDIX A

CIRCULANT AND BLOCK-CIRCULANT MATRIX

Given a d-dimensional vector (known as a *generating kernel*)

$$\mathbf{q} = (q_{-\lfloor d/2 \rfloor}, \dots, q_{-1}, q_0, q_1, \dots, q_{\lfloor d/2 \rfloor - 1})^T,$$

a $d \times d$ circulant matrix is constructed as

$$\mathcal{C}(\mathbf{q}) = \begin{pmatrix} q_0 & q_1 & q_2 & \cdots & q_{-1} \\ q_{-1} & q_0 & q_1 & \cdots & q_{-2} \\ q_{-2} & q_{-1} & q_0 & \cdots & q_{-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_1 & q_2 & q_3 & \cdots & q_0 \end{pmatrix}.$$

The rows of $\mathcal{C}(\mathbf{q})$ are circularly shifted copies of $\mathbf{q}^T$. Multiplication of $\mathcal{C}(\mathbf{q})$ with a d-dimensional vector $\mathbf{u}$ is equivalent to convolving vectors $\mathbf{q}$ and $\mathbf{u}$, with circular (Dirichlet) boundary handling. The basis functions of the d-point DFT form a complete set of eigenvectors for any circulant matrix, regardless of the choice of generating kernel as

$$\mathcal{C}(\mathbf{q}) = F\mathcal{D}(\tilde{\mathbf{q}})F^{\dagger},$$

where F is a matrix containing the DFT basis, $F^{\dagger}$ is the complex-conjugated transpose (used to compute the forward DFT), and $\mathcal{D}(\tilde{\mathbf{q}})$ is a diagonal matrix containing the d-point DFT of $\mathbf{q}$. Thus, multiplying matrix $\mathcal{C}(q)$ with vector $\mathbf{u}$ can be computed as

$$\mathcal{C}(\mathbf{q})\mathbf{u} = F\mathcal{D}(\tilde{\mathbf{q}})F^{\dagger}\mathbf{u} = F((F^{\dagger}\mathbf{q}) \otimes (F^{\dagger}\mathbf{u})),$$

generally known as the convolution theorem. Since the DFT may be implemented with $O(d \log d)$ operations, this expression often provides an efficient implementation of convolution with $\mathbf{q}$.

Representation of 2D convolutions (e.g., for images) requires a second-order circulant (also called *block-circulant*) matrix, which can be constructed by recursively applying the circulant structure. Analogous to the circulant matrix, the 2D DFT of the shifted symmetric reflection of Q are the eigenvalues of $\mathcal{C}(Q)$ and the corresponding 2D DFT basis vectors are the eigenvectors. And again, the convolution theorem provides an efficient means of implementing matrix multiplication by the circulant matrix. For a full account of the properties and computations of circulant and block-circulant matrices, readers are referred to [47].

APPENDIX B

PARAMETER LEARNING FOR 2D hGMRF

A zero-mean 2D hGMRF $\mathbf{u}$ of dimension $N \times M$ is completely determined by the generating kernel Q of its block-circulant precision matrix. The density of $\mathbf{u}$ can be expressed as

$$p(\mathbf{u}) = \sqrt{\frac{|\mathcal{C}(Q)|}{(2\pi)^{MN}}} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathcal{C}(Q) \mathbf{u} \right),$$

where we have abused notation a bit by using $\mathbf{u}$ to represent the vectorized MRF. Assuming a rectangular Markov neighborhood of size $N' \times M'$, where $M' \ll M$ and $N' \ll N$, estimation of Q corresponds to the determination of the central $N' \times M'$ entries (all others must be zero).

Given a set of independent samples of $\mathbf{u}$, $\{\mathbf{u}_k\}_{k=1}^K$, the generating kernel Q can be estimated by maximizing the log likelihood:

$$L(Q) = \sum_{k=1}^K \log p(\mathbf{u}_k)$$
$$= \frac{K}{2} \log |\mathcal{C}(Q)| - \frac{KMN}{2} \log 2\pi - \frac{1}{2} \sum_{k=1}^K \mathbf{u}_k^T \mathcal{C}(Q) \mathbf{u}_k.$$

Using the eigendecomposition property of block-circulant matrix and neglecting the constant term, we can write a modified likelihood function as

$$\tilde{L}(Q) = \frac{K}{2} \log \left| F\mathcal{D}(\tilde{Q})F^\dagger \right| - \frac{1}{2} \sum_{k=1}^K \mathbf{u}_k^T F\mathcal{D}(\tilde{Q})F^\dagger \mathbf{u}_k$$
$$= \frac{K}{2} \log \left| \mathcal{D}(\tilde{Q}) \right| - \frac{1}{2} \sum_{k=1}^K \tilde{\mathbf{u}}_k^T \mathcal{D}(\tilde{Q}) \tilde{\mathbf{u}}_k$$
$$= \frac{K}{2} \sum_{i=1}^{MN} \log \tilde{Q}_i - \frac{1}{2} \sum_{i=1}^{MN} \tilde{Q}_i \sum_{k=1}^K (\tilde{\mathbf{u}}_k)_i^2,$$

where $\tilde{\mathbf{u}} = F^\dagger \mathbf{u}$ is the DFT of $\mathbf{u}$. The optimal Q is then obtained by maximizing $\tilde{L}(Q)$ subject to the constraint that the solution leads to a symmetric and positive definite matrix $\mathcal{C}(Q)$. It can be shown that $\tilde{L}(Q)$ is a convex function and the constraints form a convex set in the feasible space of Q. Thus, solving for the optimal Q is a convex optimization problem and there are a variety of iterative solutions that are guaranteed to converge to the global optimum [48].

APPENDIX C

SAMPLING 2D hGMRF

A sample of a zero-mean 2D hGMRF $\mathbf{u}$ of dimension $N \times M$ with precision $\mathcal{C}(Q)$ can be obtained from a sample of white Gaussian noise $\mathbf{w}$ of the same dimension, by computing

$$\mathbf{u} = F\mathcal{D}\left(\sqrt{\tilde{Q}} \right)^{-1} F^\dagger \mathbf{w}. \tag{15}$$

That is, compute the DFT of the noise, divide (element-wise) by the square root of the DFT of the desired generating kernel, $\tilde{Q}$, and then invert the DFT. A similar algorithm has been used for texture synthesis [49], as well as for Monte Carlo sampling for image restoration [50].

It is easy to verify that the resulting MRF has the desired covariance structure:

$$E(\mathbf{u}\mathbf{u}^T) = E\left[F\mathcal{D}\left(\sqrt{\tilde{Q}} \right)^{-1} F^\dagger \mathbf{w}\mathbf{w}^T F\mathcal{D}\left(\sqrt{\tilde{Q}} \right)^{-1} F^\dagger \right]$$
$$= F\mathcal{D}\left(\sqrt{\tilde{Q}} \right)^{-1} F^\dagger E(\mathbf{w}\mathbf{w}^T) F\mathcal{D}\left(\sqrt{\tilde{Q}} \right)^{-1} F^\dagger$$
$$= F\mathcal{D}(\tilde{Q})^{-1} F^\dagger = \mathcal{C}(Q)^{-1},$$

where we have used the Hermitian property of matrix F and the fact that $E(\mathbf{w}\mathbf{w}^T) = I$.

REFERENCES

[1] P. Burt and E. Adelson, "The Laplacian Pyramid as a Compact Image Code," *IEEE Trans. Comm.,* vol. 31, no. 4, pp. 532-540, 1981.
[2] D.J. Field, "Relations between the Statistics of Natural Images and the Response Properties of Cortical Cells," *J. Optical Soc. Am.,* vol. 4, no. 12, pp. 2379-2394, 1987.
[3] S.G. Mallat, "A Theory for Multiresolution Signal Decomposition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 11, pp. 674-697, 1989.
[4] J. Shapiro, "Embedded Image Coding Using Zerotrees of Wavelet Coefficients," *IEEE Trans. Signal Processing,* vol. 41, no. 12, pp. 3445-3462, Dec. 1993.
[5] E.P. Simoncelli, "Statistical Models for Images: Compression, Restoration and Synthesis," *Proc. 31st Asilomar Conf. Signals, Systems and Computers,* vol. 1, pp. 673-678, Nov. 1997.
[6] R.W. Buccigrossi and E.P. Simoncelli, "Image Compression via Joint Statistical Characterization in the Wavelet Domain," *IEEE Trans. Image Processing,* vol. 8, no. 12, pp. 1688-1701, Dec. 1999.
[7] E.P. Simoncelli and E.H. Adelson, "Noise Removal via Bayesian Wavelet Coring," *Proc. Third IEEE Int'l Conf. Image Processing,* vol. 1, pp. 379-382, Sept. 1996.
[8] A. Hyvärinen, P.O. Hoyer, and M. Inki, "Topographic ICA as a Model of Natural Image Statistics," *Proc. First IEEE Int'l Workshop Biologically Motivated Computer Vision,* 2000.
[9] J. Huang and D. Mumford, "Statistics of Natural Images and Models," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 1999.
[10] P. Gehler and M. Welling, "Products of "Edge-Perts"," *Proc. Advances in Neural Information Processing Systems,* Y. Weiss, B. Schölkopf, and J. Platt, eds., pp. 419-426, 2006.
[11] A. Srivastava, X. Liu, and U. Grenander, "Universal Analytical Forms for Modeling Image Probability," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 9, pp. 217-232, Sept. 2002.
[12] Y. Teh, M. Welling, and S. Osindero, "Energy-Based Models for Sparse Overcomplete Representations," *J. Machine Learning Research,* vol. 4, pp. 1235-1260, citeseer.ist.psu.edu/teh03energy based.html, 2003.
[13] L. Parra, C. Spence, and P. Sajda, "Higher-Order Statistical Properties Arising from the Non-Stationarity of Natural Signals," *Proc. Advances in Neural Information Processing Systems,* vol. 13, 2000.
[14] L. Sendur and I.W. Selesnick, "Bivariate Shrinkage Functions for Wavelet-Based Denoising Exploiting Interscale Dependency," *IEEE Trans. Signal Processing,* vol. 50, no. 11, pp. 2744-2756, 2002.
[15] D.F. Andrews and C.L. Mallows, "Scale Mixtures of Normal Distributions," *J. Royal Statistical Soc., Series B,* vol. 36, no. 1, pp. 99-102, 1974.
[16] M.J. Wainwright and E.P. Simoncelli, "Scale Mixtures of Gaussians and the Statistics of Natural Images," *Proc. Advances in Neural Information Processing Systems,* S.A. Solla, T.K. Leen, and K.-R. Müller, eds., vol. 12, pp. 855-861, May 2000.

[17] 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,* vol. 12, no. 11, pp. 1338-1351, Nov. 2003.

[18] J. Guerrero-Colon, L. Mancera, and J. Portilla, "Image Restoration Using Space-Variant Gaussian Scale Mixtures in Overcomplete Pyramids," *IEEE Trans. Image Processing,* vol. 17, no. 1, pp. 27-41, Jan. 2008.

[19] 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,* vol. 11, no. 1, pp. 89-123, July 2001.

[20] J. Romberg, H. Choi, and R. Baraniuk, "Bayesian Tree-Structured Image Modeling Using Wavelet-Domain Hidden Markov Models," *IEEE Trans. Image Processing,* vol. 10, no. 7, July 2001.

[21] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 6, pp. 721-741, 1984.

[22] F. Jeng and J. Woods, "Compound Gauss-Markov Random Fields for Image Estimation," *IEEE Trans. Signal Processing,* vol. 39, no. 3, pp. 683-697, 1991.

[23] S.C. Zhu, Y. Wu, and D. Mumford, "Filters, Random Fields and Maximum Entropy (FRAME): Towards a Unified Theory for Texture Modeling," *Int'l J. Computer Vision,* vol. 27, no. 2, pp. 107-126, 1998.

[24] W.T. Freeman, E.C. Pasztor, and O.T. Carmichael, "Learning Low-Level Vision," *Int'l J. Computer Vision,* vol. 40, no. 1, pp. 25-47, Oct. 2000.

[25] M. Tappen, C. Liu, E. Adelson, and W. Freeman, "Learning Gaussian Conditional Random Fields for Low-Level Vision," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1-8, 2007.

[26] P. Winkler, *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods,* second ed. Springer, 2003.

[27] S. Roth and M. Black, "Fields of Experts: A Framework for Learning Image Priors," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 860-867, citeseer.ist.psu.edu/729276.html, 2005.

[28] S. Lyu and E.P. Simoncelli, "Statistical Modeling of Images with Fields of Gaussian Scale Mixtures," *Proc. Advances in Neural Information Processing Systems,* B. Schölkopf, J. Platt, and T. Hofmann, eds., vol. 19, May 2007.

[29] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, and D.J. Heeger, "Shiftable Multi-Scale Transforms," *IEEE Trans. Information Theory,* vol. 38, no. 2, pp. 587-607, special issue on wavelets, Mar. 1992.

[30] B. Wegmann and C. Zetzsche, "Statistical Dependencies between Orientation Filter Outputs Used in Human Vision Based Image Code," *Proc. Visual Comm. and Image Processing,* vol. 1360, pp. 909-922, 1990.

[31] M. Welling, G.E. Hinton, and S. Osindero, "Learning Sparse Topographic Representations with Products of Student t-Distributions," *Proc. Advances in Neural Information Processing Systems,* pp. 1359-1366, 2002.

[32] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications.* Chapman and Hall/CRC, 2005.

[33] J. Portilla, V. Strela, M.J. Wainwright, and E.P. Simoncelli, "Adaptive Wiener Denoising Using a Gaussian Scale Mixture Model in the Wavelet Domain," *Proc. Eighth IEEE Int'l Conf. Image Processing,* vol. 2, pp. 37-40, Oct. 2001.

[34] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes,* second ed. Cambridge, 2002.

[35] M.I. Jordan, Z. Ghahramani, T. Jaakkola, and L.K. Saul, "An Introduction to Variational Methods for Graphical Models," *Machine Learning,* vol. 37, no. 2, pp. 183-233, citeseer.ist.psu.edu/jordan98introduction.html, 1999.

[36] M. Figueiredo and J. Leitão, "Unsupervised Image Restoration and Edge Location Using Compound Gauss-Markov Random Fields and MDL Principle," *IEEE Trans. Image Processing,* vol. 6, no. 8, pp. 1089-1122, 1997.

[37] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image Denoising by Sparse 3D Transform-Domain Collaborative Filtering," *IEEE Trans. Image Processing,* vol. 16, no. 6, pp. 1064-1083, 2007.

[38] J. Portilla and E.P. Simoncelli, "Image Restoration Using Gaussian Scale Mixtures in the Wavelet Domain," *Proc. 10th IEEE Int'l Conf. Image Processing,* vol. 2, pp. 965-968, Sept. 2003.

[39] M. Elad and M. Aharon, "Image Denoising via Sparse and Redundant Representations over Learned Dictionaries," *IEEE Trans. Image Processing,* vol. 15, no. 12, pp. 3736-3745, Dec. 2006.

[40] E.P. Simoncelli and E.H. Adelson, "Subband Transforms," *Subband Image Coding,* J.W. Woods, ed., chapter 4, pp. 143-192, 1990.

[41] R.R. Coifman and D.L. Donoho, "Translation-Invariant De-Noising," *Wavelets and Statistics,* A. Antoniadis and G. Oppenheim, eds., Springer-Verlag, 1995.

[42] M. Raphan and E.P. Simoncelli, "Optimal Denoising in Redundant Bases," *Proc. 14th IEEE Int'l Conf. Image Processing,* Sept. 2007.

[43] Y. Karklin and M.S. Lewicki, "A Hierarchical Bayesian Model for Learning Non-Linear Statistical Regularities in Non-Stationary Natural Signals," *Neural Computation,* vol. 17, no. 2, pp. 397-423, 2005.

[44] A. Hyvärinen, J. Hurri, and J. Väyrynen, "Bubbles: A Unifying Framework for Low-Level Statistical Properties of Natural Image Sequences," *J. Optical Soc. Am. A,* vol. 20, no. 7, pp. 1237-1252, 2003.

[45] D.K. Hammond and E.P. Simoncelli, "Image Denoising with an Orientation-Adaptive Gaussian Scale Mixture Model," *Proc. 13th IEEE Int'l Conf. Image Processing,* pp. 1433-1436, Oct. 2006.

[46] Z. Wang and E.P. Simoncelli, "Local Phase Coherence and the Perception of Blur," *Proc. Advances in Neural Information Processing Systems,* vol. 16, 2003.

[47] R.M. Gray, "Toeplitz and Circulant Matrices: A Review," *Foundations and Trends in Comm. and Information Theory,* vol. 2, no. 3, pp. 155-239, 2006.

[48] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge Univ. Press, 2005.

[49] R. Chellappa, S. Chatterjee, and R. Bagdazian, "Texture Synthesis and Compression Using Gaussian-Markov Random Field Models," *IEEE Trans. Systems, Man, and Cybernetics,* vol. 15, no. 3, pp. 298-303, Mar. 1985.

[50] D. Geman and C. Yang, "Nonlinear Image Recovery with Half-Quadratic Regularization," *IEEE Trans. Image Processing,* vol. 4, no. 7, pp. 932-946, 1995.

Siwei Lyu received the BS degree in information science and the MS degree in computer science from Peking University, Beijing, in 1997 and 2000, respectively, and the PhD degree in computer science from Dartmouth College in 2007. He is currently an assistant professor in computer science at the University at Albany, State University of New York. From 2000 to 2001, he was an assistant researcher at Microsoft Research Asia, Beijing, China. From 2005 to 2008, he was a postdoctoral research associate at New York University. His research interests include image processing and forensics, machine learning, and computer vision. He is a member of the IEEE.

Eero P. Simoncelli received the BS degree (summa cum laude) in physics from Harvard University in 1984 and the MS and PhD degrees in electrical engineering from Massachusetts Institute of Technology in 1988 and 1993, respectively. He studied applied mathematics at Cambridge University for a year and a half. He was an assistant professor in the Computer and Information Science Department at the University of Pennsylvania from 1993 to 1996. In September 1996, he moved to New York University, where he is currently an associate professor in neural science and mathematics. In August 2000, he became an associate investigator at the Howard Hughes Medical Institute under their new program in computational biology. His research interests span a wide range of topics in the representation and analysis of visual images, in both machine and biological systems. He is a fellow of the IEEE.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.