

Image Denoising Using Scale Mixtures of Gaussians in the Wavelet Domain

Javier Portilla, Vasily Strela, Martin J. Wainwright, and Eero P. Simoncelli

Abstract—We describe a method for removing noise from digital images, based on a statistical model of the coefficients of an overcomplete multiscale oriented basis. Neighborhoods of coefficients at adjacent positions and scales are modeled as the product of two independent random variables: a Gaussian vector and a hidden positive scalar multiplier. The latter modulates the local variance of the coefficients in the neighborhood, and is thus able to account for the empirically observed correlation between the coefficient amplitudes. Under this model, the Bayesian least squares estimate of each coefficient reduces to a weighted average of the local linear estimates over all possible values of the hidden multiplier variable. We demonstrate through simulations with images contaminated by additive white Gaussian noise that the performance of this method substantially surpasses that of previously published methods, both visually and in terms of mean squared error.

Index Terms—Bayesian estimation, Gaussian scale mixtures, hidden Markov model, natural images, noise removal, overcomplete representations, statistical models, steerable pyramid.

THE artifacts arising from many imaging devices are quite different from the images that they contaminate, and this difference allows humans to “see past” the artifacts to the underlying image. The goal of image restoration is to relieve human observers from this task (and perhaps even to improve upon their abilities) by reconstructing a plausible estimate of the original image from the distorted or noisy observation. A prior probability model for both the noise and for uncorrupted images is of central importance for this application.

Modeling the statistics of natural images is a challenging task, partly because of the high dimensionality of the signal. Two

basic assumptions are commonly made in order to reduce dimensionality. The first is that the probability structure may be defined *locally*. Typically, one makes a Markov assumption, that the probability density of a pixel, when conditioned on a set of neighbors, is independent of the pixels beyond the neighborhood. The second is an assumption of spatial *homogeneity*: the distribution of values in a neighborhood is the same for all such neighborhoods, regardless of absolute spatial position. The Markov random field model that results from these two assumptions is commonly simplified by assuming the distributions are Gaussian. This last assumption is problematic for image modeling, where the complexity of local structures is not well described by Gaussian densities.

The power of statistical image models can be substantially improved by transforming the signal from the pixel domain to a new representation. Over the past decade, it has become standard to initiate computer-vision and image processing tasks by decomposing the image with a set of multiscale bandpass oriented filters. This kind of representation, loosely referred to as a wavelet decomposition, is effective at decoupling the high-order statistical features of natural images. In addition, it shares some basic properties of neural responses in the primary visual cortex of mammals which are presumably adapted to efficiently represent the visually relevant features of images.

A number of researchers have developed homogeneous local probability models for images in multiscale oriented representations. Specifically, the marginal distributions of wavelet coefficients are highly kurtotic, and can be described using suitable long-tailed distributions. Recent work has investigated the dependencies between coefficients, and found that the amplitudes of coefficients of similar position, orientation and scale are highly correlated. These higher order dependencies, as well as the higher order marginal statistics, may be modeled by augmenting a simple parametric model for local dependencies (e.g., Gaussian) with a set of “hidden” random variables that govern the parameters (e.g., variance). Such hidden Markov models have become widely used, for example, in speech processing.

In this article, we develop a model for neighborhoods of oriented pyramid coefficients based on a *Gaussian scale mixture* [1]: the product of a Gaussian random vector, and an independent hidden random scalar multiplier. We have previously demonstrated that this model can account for both marginal and pairwise joint distributions of wavelet coefficients [2], [3]. Here, we develop a local denoising solution as a Bayesian least squares estimator, and demonstrate the performance of this method on images corrupted by simulated additive white Gaussian noise of known variance.

Manuscript received September 29, 2002; revised April 28, 2003. During the development of this work, V. Strela was on leave from Drexel University, and was supported by an AMS Centennial Fellowship. M. J. Wainwright was supported by a NSERC-1967 Fellowship. J. Portilla and E. P. Simoncelli were supported by an NSF CAREER grant and Alfred P. Sloan Fellowship to E. P. Simoncelli, and by the Howard Hughes Medical Institute. J. Portilla was also supported by an FPI fellowship, and subsequently by a “Ramón y Cajal” grant (both from the Spanish government). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mario A. T. Figueiredo.

J. Portilla is with the Department of Computer Science and Artificial Intelligence, Universidad de Granada, 18071 Granada, Spain (e-mail: javier@decsai.ugr.es).

V. Strela is with the Department of Mathematics and Computer Science, Drexel University, Philadelphia, PA 19104 USA (e-mail: vstrela@mcs.drexel.edu).

M. J. Wainwright is with the Electrical Engineering and Computer Science Department, University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: wainwrig@eecs.berkeley.edu).

E. P. Simoncelli is with the Center for Neural Science and the Courant Institute for Mathematical Sciences, New York University, New York, NY 10003 USA (e-mail: eero.simoncelli@nyu.edu).

Digital Object Identifier 10.1109/TIP.2003.818640

I. BACKGROUND: STATISTICAL IMAGE MODELS AND DENOISING

Contemporary models of image statistics are rooted in the television engineering of the 1950s (see [4] for review), which relied on a characterization of the autocovariance function for purposes of optimal signal representation and transmission. This work, and nearly all work since, assumes that image statistics are spatially homogeneous (i.e., strict-sense stationary). Another common assumption in image modeling is that the statistics are invariant, when suitably normalized, to changes in spatial scale. The translation- and scale-invariance assumptions, coupled with an assumption of Gaussianity, provides the baseline model found throughout the engineering literature: images are samples of a Gaussian random field, with variance falling as $f^{-\gamma}$ in the frequency domain. In the context of denoising, if one assumes the noise is additive and independent of the signal, and is also a Gaussian sample, then the optimal estimator is linear.

A. Modeling Non-Gaussian Image Properties

In recent years, models have been developed to account for non-Gaussian behaviors of image statistics. One can see from casual observation that individual images are highly inhomogeneous: they typically contain many regions that are smooth, interspersed with “features” such as contours, or surface markings. This is reflected in the observed marginal distributions of bandpass filter responses, which show a large peak at zero, and tails that fall significantly slower than a Gaussian of the same variance [5]–[7] [see Fig. 1(a)]. When one seeks a linear transformation that maximizes the non-Gaussianity¹ of the marginal responses, the result is a basis set of bandpass oriented filters of different sizes spanning roughly an octave in bandwidth, e.g., [8], [9].

Due to the combination of these qualitative properties, as well as an elegant mathematical framework, multiscale oriented subband decompositions have emerged as the representations of choice for many image processing applications. Within the subbands of these representations, the kurtotic behaviors of coefficients allow one to remove noise using a point nonlinearity. Such approaches have become quite popular in the image denoising literature, and typically are chosen to perform a type of thresholding operation, suppressing low-amplitude values while retaining high-amplitude values. The concept was developed originally in the television engineering literature (where it is known as “coring,” e.g., [10]), and specific shrinkage functions have been derived under a variety of formulations, including minimax optimality under a smoothness condition [11], [12], [57], and Bayesian estimation with non-Gaussian priors, e.g., [13]–[19], [58].

In addition to the non-Gaussian marginal behavior, the responses of bandpass filters exhibit important non-Gaussian *joint* statistical behavior. In particular, even when they are second-order decorrelated, the coefficients corresponding to pairs of basis functions of similar position, orientation and scale exhibit striking dependencies [20], [21]. Casual observation indicates that large-amplitude coefficients are sparsely distributed

¹Different authors have used different measures of non-Gaussianity, but have obtained similar results.

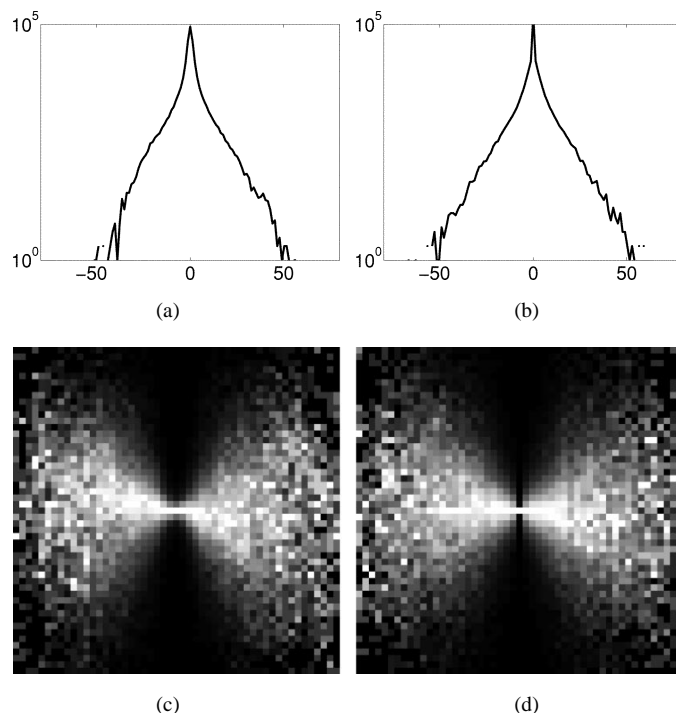


Fig. 1. Comparison of coefficient statistics from an example image subband (a vertical subband of the *Boats* image, left panels) with those arising from simulation of a local GSM model (right panels). Model parameters (covariance matrix and the multiplier prior density) are estimated by maximizing the likelihood of the observed set of wavelet coefficients. (a,b) Log marginal histograms. (c,d) Conditional histograms of two spatially adjacent coefficients. Brightness corresponds to probability, except that each column has been independently rescaled to fill the range of display intensities.

throughout the image, and tend to occur in clusters. The conditional histograms of pairs of coefficients indicates that the standard deviation of a coefficient scales roughly linearly with the amplitude of nearby coefficients [2], [21], [22] [see Fig. 1(c)].

The dependency between local coefficient amplitudes, as well as the associated marginal behaviors, can be modeled using a random field with a spatially fluctuating variance. A particularly useful example arises from the product of a Gaussian vector and a hidden scalar multiplier, known as a *Gaussian scale mixture* [1] (GSM). GSM distributions represent an important subset of the *elliptically symmetric distributions*, which are those that can be defined as functions of a quadratic norm of the random vector. Embedded in a random field, these kinds of models have been found useful in the speech-processing community [23]. A related set of models, known as autoregressive conditional heteroskedastic (ARCH) models, e.g., [24], have proven useful for many real signals that suffer from abrupt fluctuations, followed by relative “calm” periods (stock market prices, for example). These kinds of ideas have also been found effective in describing visual images. For example, Baraniuk and colleagues used a 2-state hidden multiplier variable to characterize the two modes of behavior corresponding to smooth or low-contrast textured regions and features [25], [26]. Our own work, as well as that of others, assumes that the local variance is governed by a continuous multiplier variable [2], [3], [27], [28]. This model can capture the strongly leptokurtotic behavior of the marginal densities of natural image wavelet coefficients, as well as the correlation in their local amplitudes, as illustrated in Fig. 1.

B. Empirical Bayes Denoising Using Variance-Adaptive Models

More than 20 years ago, Lee [29] suggested a two-step procedure for image denoising, in which one first estimates the local signal variance from a neighborhood of observed pixels, and then (proceeding as if this were the true variance) applies the standard linear least squares (LLS) solution. This method is a type of *empirical Bayes* estimator [30], in that a parameter of the local model is first estimated from the data, and this estimate is subsequently used to estimate the signal. This two-step denoising solution can be applied to any of the variance-adaptive models described in Section I-A, and is substantially more powerful when applied in a multiscale oriented representation. Specifically, a number of authors have estimated the local variance from a collection of wavelet coefficients at nearby positions, scales, and/or orientations, and then used these estimated variances in order to denoise the coefficients [16], [21], [28], [31]–[33].

Solutions based on GSM models, with different prior assumptions about the hidden variables, have produced some of the most effective methods for removing homogeneous additive noise from natural images to date. Our initial work in this area developed a maximum likelihood (ML) estimator [34]. Mihçak *et al.* used a maximum a posteriori (MAP) estimator based on an exponential marginal prior [28], as did Li and Orchard [35], whereas Portilla *et al.* used a lognormal prior [36]. Wainwright *et al.* developed a tree-structured Markov model to provide a global description for the set of multiplier variables [3]. Despite these successes, the two-step empirical Bayes approach is suboptimal, even when the local variance estimator is optimal, because the second step does not take into account the uncertainty associated with the variance estimated in the first step. In this paper, we derive a least squares optimal single-step Bayesian estimator.

II. IMAGE PROBABILITY MODEL

As described in Section I, multiscale representations provide a useful front-end for representing the structures of visual images. But the widely used orthonormal or biorthogonal wavelet representations are problematic for many applications, including denoising. Specifically, they are critically sampled (the number of coefficients is equal to the number of image pixels), and this constraint leads to disturbing visual artifacts (i.e., “aliasing” or “ringing”). A widely followed solution to this problem is to use basis functions designed for orthogonal or biorthogonal systems, but to reduce or eliminate the decimation of the subbands, e.g., [37].

Once the constraint of critical sampling has been dropped, however, there is no need to limit oneself to these basis functions. Significant improvement comes from the use of representations with a higher degree of redundancy, as well as increased selectivity in orientation [16], [19], [34], [38]. For the current paper, we have used a particular variant of an overcomplete tight frame representation known as a *steerable pyramid* [38], [39]. The basis functions of this multiscale 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. Other authors have developed representations with similar properties [19], [40]–[42]. Details of the steerable pyramid representation are provided in Appendix A.

A. Gaussian Scale Mixtures

Consider an image decomposed into oriented subbands at multiple scales. We denote as $x_c^{s,o}(n,m)$ the coefficient corresponding to a linear basis function at scale s , orientation o , and centered at spatial location $(2^s n, 2^s m)$. We denote as $\mathbf{x}^{s,o}(n,m)$ a *neighborhood* of coefficients clustered around this *reference coefficient*². In general, the neighborhood may include coefficients from other subbands (i.e., corresponding to basis functions at nearby scales and orientations), as well as from the same subband. In our case, we use a neighborhood of coefficients drawn from two subbands at adjacent scales, thus taking advantage of the strong statistical coupling observed through scale in multiscale representations. Details are provided in Section IV.

We assume the coefficients within each local neighborhood around a reference coefficient of a pyramid subband are characterized by a Gaussian scale mixture (GSM) model. Formally, a random vector \mathbf{x} is a Gaussian scale mixture [1] if and only if it can be expressed as the product of a zero-mean Gaussian vector \mathbf{u} and an independent positive scalar random variable \sqrt{z}

$$\mathbf{x} \stackrel{d}{=} \sqrt{z} \mathbf{u} \quad (1)$$

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

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

where N is the dimensionality of \mathbf{x} and \mathbf{u} (in our case, the size of the neighborhood). Without loss of generality, one can assume $\mathbb{E}\{z\} = 1$, which implies $\mathbf{C}_{\mathbf{x}} = \mathbf{C}_{\mathbf{u}}$.

The conditions under which a random vector may be represented using a GSM have been studied [1]. The GSM family includes a variety of well-known families of random variables such as the α -stable family (including the Cauchy distribution), the generalized Gaussian (or stretched exponential) family and the symmetrized Gamma family [3]. GSM densities are symmetric and zero-mean, and they have leptokurtotic marginal densities (i.e., heavier tails than a Gaussian). A key property of the GSM model is that the density of \mathbf{x} is Gaussian when conditioned on z . Also, the normalized vector \mathbf{x}/\sqrt{z} is Gaussian.

²For notational simplicity, we drop the superscripts s, o and indices (n, m) in the following development.

B. GSM Model for the Wavelet Coefficients

As explained in Section I and illustrated in Fig. 1, a GSM model can account for both the shape of wavelet coefficient marginals and the strong correlation between the amplitudes of neighbor coefficients [2], [3]. In order to construct a global model for images from this local description, one must specify both the neighborhood structure of the coefficients, and the distribution of the multipliers. The definition of (and calculations using) the global model is considerably simplified by partitioning the coefficients into nonoverlapping neighborhoods. One can then specify either a marginal model for the multipliers (treating them as independent variables) [43], or specify a joint density over the full set of multipliers [3]. Unfortunately, the use of disjoint neighborhoods leads to noticeable denoising artifacts at the discontinuities introduced by the neighborhood boundaries.

An alternative approach is to use a GSM as a local description of the behavior of the cluster of coefficients centered at each coefficient in the pyramid. Since the neighborhoods overlap, each coefficient will be a member of many neighborhoods. The local model implicitly defines a global (Markov) model, described by the conditional density of a coefficient in the cluster given its surrounding neighborhood, assuming conditional independence on the rest of the coefficients. But the structure of the resulting model is such that performing statistical inference (i.e., computing Bayes estimates) in an exact way is quite challenging. In this paper, we simply solve the estimation problem for the reference coefficient at the center of each neighborhood independently.

C. Prior Density for Multiplier

To complete the model, we need to specify the probability density, $p_z(z)$, of the multiplier. Several authors have suggested the generalized Gaussian (stretched exponential) family of densities as an appropriate description of wavelet coefficient marginal densities [7], [13], [17]: $p_x(x) \propto \exp(-|x/s|^p)$, where the scaling variable s controls the width of the distribution, and the exponent p controls the shape (in particular, the heaviness of the tails), and is typically estimated to lie in the range [0.5, 0.8] for image subbands. Although these can be expressed as GSM's, the density of the associated multiplier has no closed form expression, and thus this solution is difficult to implement.

In previous work [36], we noted that for the case $N = 1$, the density of the log coefficient magnitude, $\log|x|$, may be written as a convolution of the densities of $\log|u|$ and $\log\sqrt{z}$. Since the density of u is known, this means that estimation of the density of $\log\sqrt{z}$ may be framed as a deconvolution problem. The resulting estimated density may be approximated by a Gaussian, corresponding to a lognormal prior for the z . This solution has two important drawbacks. First, it is only extrapolable to the $N > 1$ case when all the neighbors have the same marginal statistics, which, in practice requires they all belong to the same subband. Second, it is estimated from the noise-free coefficients, and it is difficult to extend it for use in the noisy case.

We have also investigated a more direct maximum likelihood approach for estimating a nonparametric $p_z(z)$ from an observed set of neighborhood vectors

$$\hat{p}_z(z) = \arg \max_{p_z(z)} \sum_{m=1}^M \log \left(\int_0^\infty p(\mathbf{x}_m|z) p_z(z) dz \right) \quad (3)$$

where the sum is over the neighborhoods. Note that the estimate, $\hat{p}_z(z)$, must be constrained to positive values, and must have unit area. We have developed an efficient algorithm for computing this solution numerically. One advantage of the ML solution is that it is easily extended for use with the noisy observations, by replacing \mathbf{x}_m with the noisy observation.

A fourth choice is a so-called *noninformative prior* [44], which has the advantage that it does not require the fitting of any parameters to the noisy observation. Such solutions have been used in establishing marginal priors for image denoising [45]. We have examined the most widely used solution, known as *Jeffrey's prior* (see [44]). In the context of estimating the multiplier z from coefficients \mathbf{x} , this takes the form:

$$p_z(z) \propto \sqrt{I(z)}, \quad I(z) = \mathbb{E} \left\{ -\frac{\partial^2 \log p(\mathbf{x}|z)}{\partial z^2} \right\}$$

where $I(z)$ is the Fisher information matrix. Computing this for the GSM model is straightforward

$$\begin{aligned} -\frac{\partial^2 \log p(\mathbf{x}|z)}{\partial z^2} &= \frac{\partial^2}{\partial z^2} \left[\frac{1}{2} \left(N \log(z) + \log |\mathbf{C}_u| \right. \right. \\ &\quad \left. \left. + \frac{\mathbf{x}^T \mathbf{C}_u^{-1} \mathbf{x}}{z} \right) \right] \\ &= \frac{N}{2z^2} + \frac{\mathbf{x}^T \mathbf{C}_u^{-1} \mathbf{x}}{2z^3}. \end{aligned}$$

Taking the square root of the expectation, and using the fact that $\mathbb{E}\{\mathbf{x}^T \mathbf{C}_u^{-1} \mathbf{x}\} = z$ we obtain Jeffrey's prior

$$p_z(z) \propto \frac{1}{z} \quad (4)$$

which corresponds to a constant prior on $\log(z)$. Note that this is an improper probability density. Nevertheless it is common to ignore this fact as long as it does not create computational problems at the estimation stage. In our case, we have set the prior to zero in the interval $[0, z_{\min})$ to prevent such problems, where z_{\min} is a small positive constant (see Section IV for details).

Of the four alternatives described above, we have found (as expected) that the ML-estimated nonparametric prior produces the best results for denoising the pyramid coefficients. But a least squares optimal estimate for the pyramid coefficients does *not* necessarily lead to a least-squares optimal estimate for the image pixels, since the pyramid representation is overcomplete. We were surprised to find that the noninformative prior typically leads to better denoising performance in the image domain (roughly +0.15 dB, on average). Given that it is also simpler and more efficient to implement, we have used it for all of the results shown in Sections III–V.

III. IMAGE DENOISING

Our procedure for image denoising uses the same top-level structure as most previously published approaches: 1) decompose the image into pyramid subbands at different scales and orientations; 2) denoise each subband, except for the lowpass residual band; and 3) invert the pyramid transform, obtaining the denoised image. We assume the image is corrupted by independent additive white Gaussian noise of known variance (note that the method can also handle nonwhite Gaussian noise of known covariance). A vector \mathbf{y} corresponding to a neighborhood of N observed coefficients of the pyramid representation can be expressed as

$$\mathbf{y} = \mathbf{x} + \mathbf{w} = \sqrt{z}\mathbf{u} + \mathbf{w}. \quad (5)$$

Note that the assumed GSM structure of the coefficients, coupled with the assumption of independent additive Gaussian noise, means that the three random variables on the right side of (5) are independent.

Both \mathbf{u} and \mathbf{w} are zero-mean Gaussian vectors, with associated covariance matrices \mathbf{C}_u and \mathbf{C}_w . The density of the observed neighborhood vector conditioned on z is a zero-mean Gaussian, with covariance $\mathbf{C}_{y|z} = z\mathbf{C}_u + \mathbf{C}_w$

$$p(\mathbf{y}|z) = \frac{\exp\left(\frac{-\mathbf{y}^T(z\mathbf{C}_u + \mathbf{C}_w)^{-1}\mathbf{y}}{2}\right)}{\sqrt{(2\pi)^N |z\mathbf{C}_u + \mathbf{C}_w|}}. \quad (6)$$

The neighborhood noise covariance, \mathbf{C}_w , is obtained by decomposing a delta function $\sigma\sqrt{N_y N_x}\delta(n, m)$ into pyramid subbands, where (N_y, N_x) are the image dimensions. This signal has the same power spectrum as the noise, but it is free from random fluctuations. Elements of \mathbf{C}_w may then be computed directly as sample covariances (i.e., by averaging the products of pairs of coefficients over all the neighborhoods of the subband). This procedure is easily generalized for nonwhite noise, by replacing the delta function with the inverse Fourier transform of the square root of the noise power spectral density. Note that the entire procedure may be performed off-line, as it is signal-independent.

Given \mathbf{C}_w , the signal covariance \mathbf{C}_u can be computed from the observation covariance matrix \mathbf{C}_y . We compute \mathbf{C}_y from $\mathbf{C}_{y|z}$ by taking expectations over z :

$$\mathbf{C}_y = \mathbb{E}\{z\}\mathbf{C}_u + \mathbf{C}_w.$$

Without loss of generality, we set $\mathbb{E}\{z\} = 1$, resulting in:

$$\mathbf{C}_u = \mathbf{C}_y - \mathbf{C}_w. \quad (7)$$

We force \mathbf{C}_u to be positive semidefinite by performing an eigenvector decomposition and setting any possible negative eigenvalues (nonexisting or negligible, in most cases) to zero.

A. Bayes Least Squares Estimator

For each neighborhood, we wish to estimate x_c , the reference coefficient at the center of the neighborhood, from \mathbf{y} , the set of

observed (noisy) coefficients. The Bayes least squares (BLS) estimate is just the conditional mean

$$\begin{aligned} \mathbb{E}\{x_c|\mathbf{y}\} &= \int x_c p(x_c|\mathbf{y}) dx_c \\ &= \int \int_0^\infty x_c p(x_c, z|\mathbf{y}) dz dx_c \\ &= \int \int_0^\infty x_c p(x_c|\mathbf{y}, z) p(z|\mathbf{y}) dz dx_c \\ &= \int_0^\infty p(z|\mathbf{y}) \mathbb{E}\{x_c|\mathbf{y}, z\} dz \end{aligned} \quad (8)$$

where we have assumed uniform convergence in order to exchange the order of integration. Thus, the solution is the average of the Bayes least squares estimate of x when conditioned on z , weighted by the posterior density, $p(z|\mathbf{y})$. We now describe each of these individual components.

B. Local Wiener Estimate

The key advantage of the GSM model is that the coefficient neighborhood vector \mathbf{x} is Gaussian when conditioned on z . This fact, coupled with the assumption of additive Gaussian noise means that the expected value inside the integral of (8) is simply a local linear (Wiener) estimate. Writing this for the full neighborhood vector

$$\mathbb{E}\{\mathbf{x}|\mathbf{y}, z\} = z\mathbf{C}_u(z\mathbf{C}_u + \mathbf{C}_w)^{-1}\mathbf{y}. \quad (9)$$

We can simplify the dependence of this expression on z by diagonalizing the matrix $z\mathbf{C}_u + \mathbf{C}_w$. Specifically, let \mathbf{S} be the symmetric square root of the positive definite matrix \mathbf{C}_w (i.e., $\mathbf{C}_w = \mathbf{S}\mathbf{S}^T$), and let $\{\mathbf{Q}, \mathbf{\Lambda}\}$ be the eigenvector/eigenvalue expansion of the matrix $\mathbf{S}^{-1}\mathbf{C}_u\mathbf{S}^{-T}$. Then

$$\begin{aligned} z\mathbf{C}_u + \mathbf{C}_w &= z\mathbf{C}_u + \mathbf{S}\mathbf{S}^T \\ &= \mathbf{S}(z\mathbf{S}^{-1}\mathbf{C}_u\mathbf{S}^{-T} + \mathbf{I})\mathbf{S}^T \\ &= \mathbf{S}\mathbf{Q}(z\mathbf{\Lambda} + \mathbf{I})\mathbf{Q}^T\mathbf{S}^T. \end{aligned} \quad (10)$$

Note this diagonalization does not depend on z , and thus need only be computed once for each subband. We can now simplify (9) as follows:

$$\begin{aligned} \mathbb{E}\{\mathbf{x}|\mathbf{y}, z\} &= z\mathbf{C}_u\mathbf{S}^{-T}\mathbf{Q}(z\mathbf{\Lambda} + \mathbf{I})^{-1}\mathbf{Q}^T\mathbf{S}^{-1}\mathbf{y} \\ &= z\mathbf{S}\mathbf{S}^{-1}\mathbf{C}_u\mathbf{S}^{-T}\mathbf{Q}(z\mathbf{\Lambda} + \mathbf{I})^{-1}\mathbf{Q}^T\mathbf{S}^{-1}\mathbf{y} \\ &= z\mathbf{S}\mathbf{Q}\mathbf{\Lambda}(z\mathbf{\Lambda} + \mathbf{I})^{-1}\mathbf{Q}^T\mathbf{S}^{-1}\mathbf{y} \\ &= z\mathbf{M}\mathbf{\Lambda}(z\mathbf{\Lambda} + \mathbf{I})^{-1}\mathbf{v} \end{aligned} \quad (11)$$

where $\mathbf{M} = \mathbf{S}\mathbf{Q}$, and $\mathbf{v} = \mathbf{M}^{-1}\mathbf{y}$. Finally, we restrict the estimate to the reference coefficient, as needed for the solution of (8)

$$\mathbb{E}\{x_c|\mathbf{y}, z\} = \sum_{n=1}^N \frac{zm_{cn}\lambda_n v_n}{z\lambda_n + 1} \quad (12)$$

where m_{ij} represents an element (i -th row, j -th column) of the matrix \mathbf{M} , λ_n are the diagonal elements of $\mathbf{\Lambda}$, v_n the elements of \mathbf{v} , and c is the index of the reference coefficient within the neighborhood vector.

C. Posterior Distribution of the Multiplier

The other component of the solution given in (8) is the distribution of the multiplier, conditioned on the observed neighborhood values. We use Bayes' rule to compute this

$$p(z|\mathbf{y}) = \frac{p(\mathbf{y}|z)p_z(z)}{\int_0^\infty p(\mathbf{y}|\alpha)p_z(\alpha) d\alpha}. \quad (13)$$

As discussed in Section II-C, we choose a noninformative Jeffrey's prior, corrected at the origin, for the function $p_z(z)$. The conditional density $p(\mathbf{y}|z)$ is given in (6), and its computation may be simplified using the relationship in (10) and the definition of \mathbf{v}

$$p(\mathbf{y}|z) = \frac{\exp\left(-\frac{1}{2} \sum_{n=1}^N \frac{v_n^2}{z\lambda_n+1}\right)}{\sqrt{(2\pi)^N |\mathbf{C}_w| \prod_{n=1}^N (z\lambda_n+1)}}. \quad (14)$$

Summarizing our denoising algorithm

- 1) Decompose the image into subbands.
- 2) For each subband (except the lowpass residual):
 - a) Compute neighborhood noise covariance, \mathbf{C}_w , from the image-domain noise covariance.
 - b) Estimate noisy neighborhood covariance, \mathbf{C}_y .
 - c) Estimate \mathbf{C}_u from \mathbf{C}_w and \mathbf{C}_y using (7).
 - d) Compute \mathbf{A} and \mathbf{M} (Section III-B).
 - e) For each neighborhood:
 - i) For each value z in the integration range:
 - A) Compute $\mathbb{E}\{x_c|\mathbf{y}, z\}$ using (12).
 - B) Compute $p(\mathbf{y}|z)$ using (14).
 - ii) Compute $p(z|\mathbf{y})$ using (13) and (4).
 - iii) Compute $\mathbb{E}\{x_c|\mathbf{y}\}$ numerically using (8).
- 3) Reconstruct the denoised image from the processed subbands and the lowpass residual.

IV. IMPLEMENTATION

We decompose the image into subbands using a specialized variant of the steerable pyramid. The representation consists of oriented bandpass bands at 8 orientations and 5 scales, 8 oriented highpass residual subbands, and one lowpass (nonoriented) residual band, for a total of 49 subbands. A detailed description of the decomposition is given in Appendix A.

We have hand-optimized the neighborhood structure (i.e., choice of spatial positions, scales and orientations). A 3×3

region surrounding the reference coefficient, together with the coefficient at the same location and orientation at the next coarser scale (the *parent*), maximizes the denoising performance, on average. Inclusion of parent coefficient has been found to provide a significant improvement in performance in a number of applications, e.g., [21], [22], [25], [26], [46]. Note that since the parent subband is sampled at half the density of the reference subband, it must be upsampled and interpolated in order to obtain values for neighborhoods at every choice of reference coefficient. Two exceptions must be applied: 1) the highpass oriented subbands, whose parents have the same number of samples as them (no interpolation is required for those parents); and 2) the subbands at the coarsest scale, which have no parent subband (we simply use the 3×3 spatial neighborhood for those subbands). Note that in terms of image pixels, the spatial extent of the neighborhood depends on the scale of the subband (the basis functions grow in size as 2^s) as is appropriate under the assumption that image statistics are scale-invariant [47], [48].

In our implementation, the integral of (8) is computed numerically. The range and sample spacing for this integration are chosen as a compromise between accuracy and computational cost. Specifically, we sample z with logarithmically uniform spacing, which we have observed to require fewer samples, for the same quality, than linear sampling. Note also that Jeffrey's improper prior for z is a constant under a logarithmic representation. We use only $S_z = 13$ samples of $\log(z)$ over an interval $[\log(z_{\min}), \log(z_{\max})]$ using steps of size 2. We have chosen $\log(z_{\min}) = -20.5$ and $\log(z_{\max}) = 3.5$. The value z_{\max} is chosen as the minimal value that guarantees in practice that the right-tails of all the posteriors are properly covered by the integration interval. In contrast, z_{\min} plays the role of ensuring that the left tail of the posterior is integrable. We have hand-optimized z_{\min} to maximize the performance of the algorithm, and have found that denoising performance is relatively insensitive to changes in this parameter. Only slightly worse results ($\simeq -0.01$ to -0.05 dB) result from choosing $\log(z_{\min})$ within the interval $[-40, -10]$, and reasonable performance ($\simeq -0.1$ to -0.2 dB) is obtained with values as low as -200 (which corresponds to $z_{\min} \simeq 10^{-37}$).

The computational cost of the pyramid transform (both forward and inverse) scales as $I_x I_y \log_2(I_x I_y)$, where $I_{\{x,y\}}$ are the dimensions of the image. The computational cost of the estimation procedure scales as $(I_x + (N_x + B_x)/2)(I_y + (N_y + B_y)/2)NK S_z$, where $N_{\{x,y\}}$ are the dimensions of the spatial subband neighborhood (3 in our case), $B_{\{x,y\}}$ the dimensions of the bandpass convolution kernels (roughly 9 in our implementation), N the full size of the neighborhood (10 in our case), K the number of orientations, and S_z the number of samples used for the distributions over z . The terms added to the image dimensions correspond to the padded boundary region that must be estimated in order to properly reconstruct the image. As a guide, running times in our current unoptimized Matlab implementation, on a Linux workstation with 1.7 GHz Intel Pentium-III CPU, are roughly 40 seconds for 256×256 images. Finally, the primary memory cost is due to storage of the pyramid coefficients (roughly $7KN_x N_y/3$ floating point numbers).

TABLE I
 DENOISING PERFORMANCE EXPRESSED AS PEAK SIGNAL-TO-NOISE RATIO, $20 \log_{10}(255/\sigma_e)$ IN dB, WHERE σ_e IS THE ERROR STANDARD DEVIATION. EVERY ENTRY IS THE AVERAGE USING EIGHT DIFFERENT NOISE SAMPLES. LAST COLUMN SHOWS THE ESTIMATED STANDARD DEVIATION OF THESE RESULTS FOR EACH NOISE LEVEL

σ / PSNR	<i>Lena</i>	<i>Barb</i>	<i>Boats</i>	<i>Fgrpt</i>	<i>House</i>	<i>Peprs</i>	σ_{PSNR}
1 / 48.13	48.46	48.37	48.44	48.46	48.85	48.38	0.009
2 / 42.11	43.23	43.29	42.99	43.05	44.07	43.00	0.012
5 / 34.15	38.49	37.79	36.97	36.68	38.65	37.31	0.014
10 / 28.13	35.61	34.03	33.58	32.45	35.35	33.77	0.017
15 / 24.61	33.90	31.86	31.70	30.14	33.64	31.74	0.024
20 / 22.11	32.66	30.32	30.38	28.60	32.39	30.31	0.031
25 / 20.17	31.69	29.13	29.37	27.45	31.40	29.21	0.037
50 / 14.15	28.61	25.48	26.38	24.16	28.26	25.90	0.049
75 / 10.63	26.84	23.65	24.79	22.40	26.41	24.00	0.061
100 / 8.13	25.64	22.61	23.75	21.22	25.11	22.66	0.070

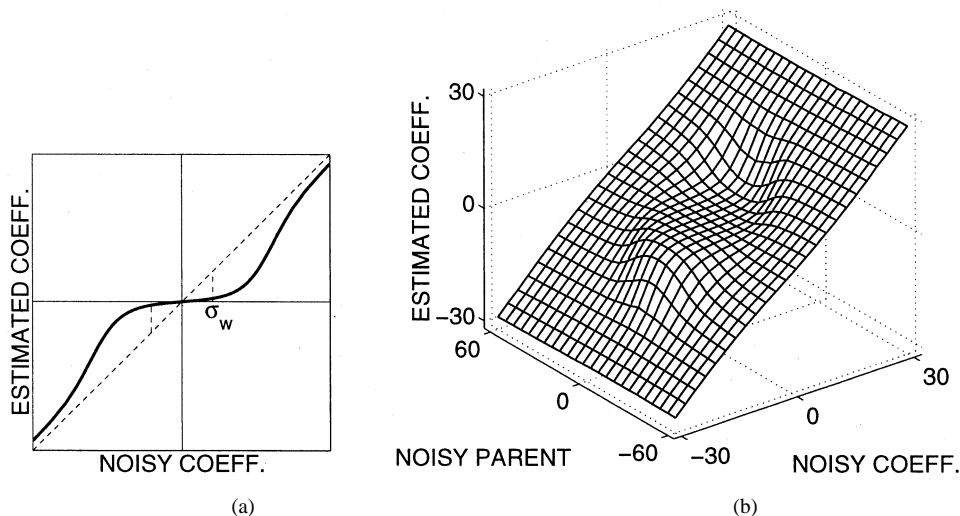


Fig. 2. Nonlinear estimation functions resulting from restriction of our method to smaller neighborhoods. (a) Neighborhood of size one (reference coefficient only) and (b) neighborhood of size two (reference coefficient plus parent).

V. RESULTS

We have tested our method on a set of 8-bit grayscale test images, of size 512×512 and 256×256 pixels, each contaminated with computer-generated additive Gaussian white noise at 10 different variances. Further information about the images is provided in Appendix B. Table I shows error variances of the denoised images, expressed as peak signal-to-noise ratios (PSNR) in decibels, for the full range of input noise levels. Note that for all images, there is very little improvement at the lowest noise level. This makes sense, since the “clean” images in fact include quantization errors, and have an implicit PSNR of 58.9 dB. At the other extreme, improvement is substantial (roughly 17 dB in the best cases).

A. Comparison to Model Variants

In order to understand the relative contribution of various aspects of our method, we considered two restricted versions of our model that are representative of the two primary denoising concepts found in the literature. The first is a Gaussian model, arising from the restriction of our model to a prior density $p_z(z)$ which is a delta function concentrated at one. This model is not

variance-adaptive, and, thus, is globally Gaussian. Note, though, that the signal covariance is modeled only locally (over the extent of the neighborhood) for each pyramid subband. As such, this denoising solution may be viewed as a regularized version of the classical linear (Wiener filter) solution. In order to implement this, we simply estimate each coefficient using (12), with z set to one.

The second restricted form of our model uses a neighborhood containing only the reference coefficient (i.e., 1×1). Under these conditions, the model describes only the marginal density of the coefficients, and the estimator reduces to application of a scalar function to the observed noisy coefficients. The function resulting from the reduction of our model to a single-element neighborhood is shown in Fig. 2(a). This is similar to the BLS solutions derived in [13], [16] for a generalized Gaussian prior, except that it is independent of the clean signal statistics, and its normalized form $\hat{x}(y)/y$ scales with the noise standard deviation of the subband, σ_w (as in [11]).

It is also instructive to examine the nonlinear estimator associated with the case of two neighbors. Fig. 2(b) shows the estimator obtained as a function of the reference coefficient and

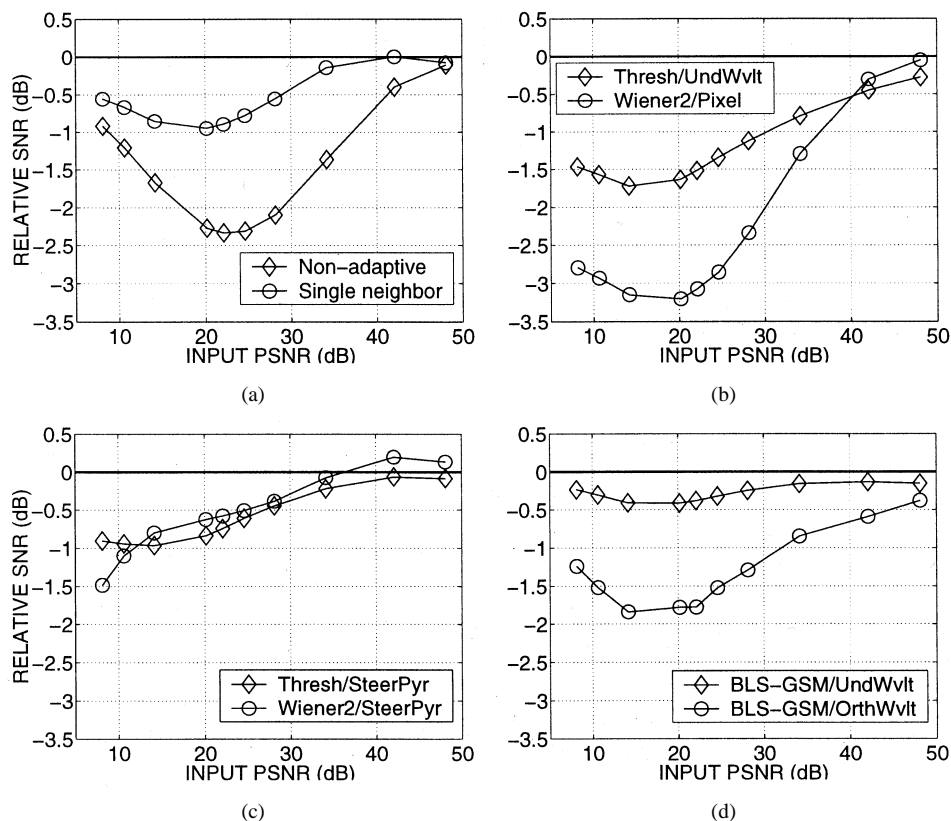


Fig. 3. Performance of other denoising methods relative to our method. Curves depict PSNR differences (in dB), averaged over three representative images (*Lena*, *Barbara*, and *Boats*) as a function of input PSNR. (a) Comparison to two restricted cases: A nonadaptive (globally Gaussian) GSM model resulting from using $p_z(z) = \delta(z - 1)$ (diamonds), and a GSM model with a neighborhood of size one (circles). (b) Comparison to hard-thresholding in an undecimated (minimum-phase, Daubechies 8-tap, 5 scales) wavelet decomposition (*diamonds*) [37], and a local variance-adaptive method in the image domain (*circles*) [29], as implemented by Matlab’s *wiener2* function. Parameters for both methods have been optimized for each image and noise level: a threshold level for the first method, and a neighborhood size (ranging from 3 to 11) for the second. (c) Two denoising algorithms applied to our steerable pyramid representation: adaptive Wiener [29], using a hand-optimized size of neighborhood (19×19 for all the images and noise levels) (circles), and hard-thresholding, optimizing the threshold for every image and noise level (diamonds). (d) Application of our BLS-GSM estimation method to coefficients of two different representations: an undecimated minimum-phase Daubechies 8-tap wavelet, using 5 scales (diamonds), and the same decomposition in its original decimated version (circles).

a coarse-scale (parent) coefficient. Loosely speaking, the reference coefficient is suppressed only when both its own amplitude and the parent’s amplitude are small. Adjacent neighbors in the same subband have a similar effect on the estimation. Sendur and Selesnick have recently developed a MAP estimator based on a circular-symmetric Laplacian density model for a coefficient and its parent [49], [50]. Their resulting shrinkage function is qualitatively similar to that of Fig. 2(b), except that ours is smoother and, due to covariance adaptation, its “dead zone” is not necessary aligned with the input axes.

Fig. 3(a) shows a comparison of our full model and the two reduced forms explained above. Note that the 1-D shrinkage solution outperforms the jointly Gaussian (nonadaptive) solution, which still provides relatively good results, especially at low SNR rates. The full model (adaptive and context-sensitive) incorporates the advantages of the two subcases, and thus outperforms both of them.

We have also examined the relative importance of other aspects of our method. Table II shows the decrease in PSNR that results when each of a set of features is removed. The first three columns correspond to features of the representation, the next two to features of the model, and the last to the estimation method. Within the first group, decreasing the number of orientation bands from $K = 8$ to $K = 4$ (*Orid*) leads to a significant

drop in performance. We have also found that further increasing the number of orientations leads to additional PSNR improvement, at the expense of considerable computational and storage cost. The second column (*OrHPR*) shows the effect of not partitioning the highpass residual band into oriented components (the standard form of the pyramid, as used in our previous denoising work [34], [36], has only a single nonoriented highpass residual band). The third column (*Bdry*) shows the reduction in performance that results when switching from mirror-reflected extension to periodic boundary handling.

The first feature of the model we examined is the inclusion of the coarse-scale parent coefficient in the neighborhood. The fourth column (*Prnt*) shows that eliminating the coarse-scale parent from the neighborhood decreases performance significantly only at high noise levels. This should not be taken to mean that the parent coefficient does not provide information about the reference coefficient, but that the information is somewhat redundant with that provided by the other neighbors [22]. The next column (*Cov*), demonstrates the result of assuming uncorrelated Gaussian vectors in describing both noise and signal. The coefficients in our representation are strongly correlated, both because of inherent spectral features of the image and because of the redundancy induced by the overcomplete representation, and ignoring this correlation in the model leads to a

TABLE II
REDUCTION IN DENOISING PERFORMANCE (dB) RESULTING FROM REMOVAL OF MODEL COMPONENTS, SHOWN AT 3 DIFFERENT NOISE CONTAMINATION RATES. RESULTS ARE AVERAGED OVER *LENA*, *BARBARA*, AND *BOATS*. SEE TEXT FOR FURTHER INFORMATION

σ / PSNR	<i>Ori8</i>	<i>OrHPR</i>	<i>Bdry</i>	<i>Prnt</i>	<i>Cov</i>	<i>BLS</i>
10 / 28.13	0.18	0.21	0.12	-0.01	0.47	0.13
25 / 20.17	0.29	0.21	0.15	0.05	0.69	0.30
50 / 14.15	0.29	0.15	0.15	0.09	0.77	0.38

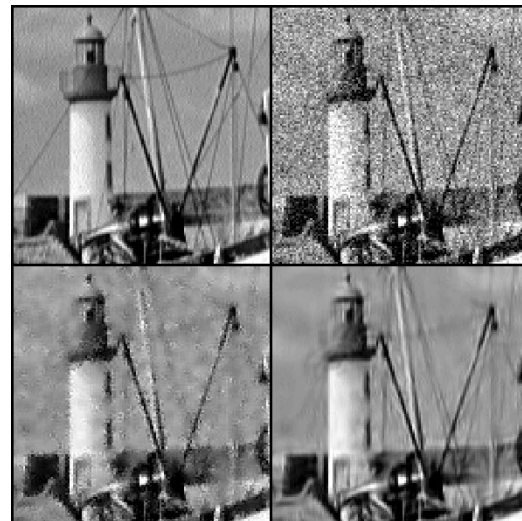
significant loss in performance. The last column (*BLS*) demonstrates a substantial reduction in performance when we replace the full *BLS* estimator with the two-step estimator (MAP estimation of the local multiplier, followed by linear estimation of the coefficient), as used in [36].

B. Comparison to Standard Methods

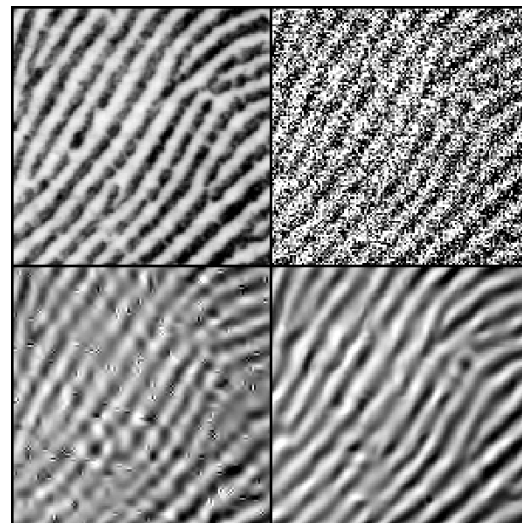
We have compared our method to two well-known and widely-available denoising algorithms: a local variance-adaptive method in the pixel domain [29] (as implemented by the Matlab function *wiener2*), and a hard thresholding method using an undecimated representation [37] with five scales based on the minimum-phase Daubechies 8-tap wavelet filter. In both cases, a single parameter (the neighborhood size or a common threshold for all the subbands) was optimized independently for each image at each noise level. Results are shown in Fig. 3(b). Our method is seen to clearly outperform the other two over the entire range of noise levels. We also see the superiority of the two multiscale methods over the pixel-domain method. Fig. 4 provides a visual comparison of example images denoised using these two algorithms. Our method produces artifacts that are significantly less visible, and at the same time is able to better preserve the features of the original image.

It is natural to ask to what extent the results in the previous comparison are due to the representation (steerable pyramid) as opposed to the estimation method itself (*BLS-GSM*). In order to answer this, we have performed two more sets of experiments, comparing the performance of different combinations of representation and estimator. First, we have applied the two estimation methods used in Fig. 3(b) to the coefficients of the steerable pyramid representation. For the adaptive Wiener method [29], we have found that the hand-optimized neighborhood size within the subbands is roughly 19×19 —much larger than in the pixel domain. For the translation invariant hard-thresholding method [37], we have optimized a common threshold for each image and noise level (note that the steerable pyramid subband impulse responses are not normalized in energy, so the common threshold needs to be properly re-scaled for each subband). Results are plotted in Fig. 3(c). It is clear that the use of the new representation improves the results and reduces the difference between the methods. The adaptive Wiener method is even seen to outperform ours at very high input SNR's. But significant differences in performance remain, and these are due entirely to the use of the *BLS-GSM* estimation method.

In a second experiment, we compared the performance of our estimation method when applied to coefficients of two dif-



(a)



(b)

Fig. 4. Comparison of denoising results on two images (cropped to 128×128 for visibility of the artifacts). (a) *Boats* image. Top-left: original image. Top-right: noisy image, $PSNR = 22.11$ dB ($\sigma = 20$). Bottom-left: denoising result using adaptive local Wiener in the image domain [29], $PSNR = 28.0$ dB. Bottom-right: our method, $PSNR = 30.4$ dB. (b) *Fingerprint* image. Top-left: original image. Top-right: noisy image, $PSNR = 8.1$ dB ($\sigma = 100$). Bottom-left: denoising result using hard thresholding in an undecimated wavelet [37] with a single optimized threshold, $PSNR = 19.0$ dB. Bottom-right: our method, $PSNR = 21.2$ dB.

ferent representations. We have chosen the most widely-used multiscale representations: the decimated and undecimated versions of a separable wavelet decomposition. In order to use the *BLS-GSM* method under an aliasing-free version of an orthogonal wavelet, we have used two fully undecimated levels for the two highest frequency scales, and have decimated by factors of two the rest of scales, producing very little aliasing and reconstruction error. This representation is analogous to the steerable pyramid: both the highpass oriented subbands and the bandpass highest frequency oriented subbands are kept at full resolution, and the rest are downsampled in a dyadic scheme. Results are plotted in Fig. 3(d), and indicate a somewhat modest decrease in performance when replacing the steerable pyramid with an

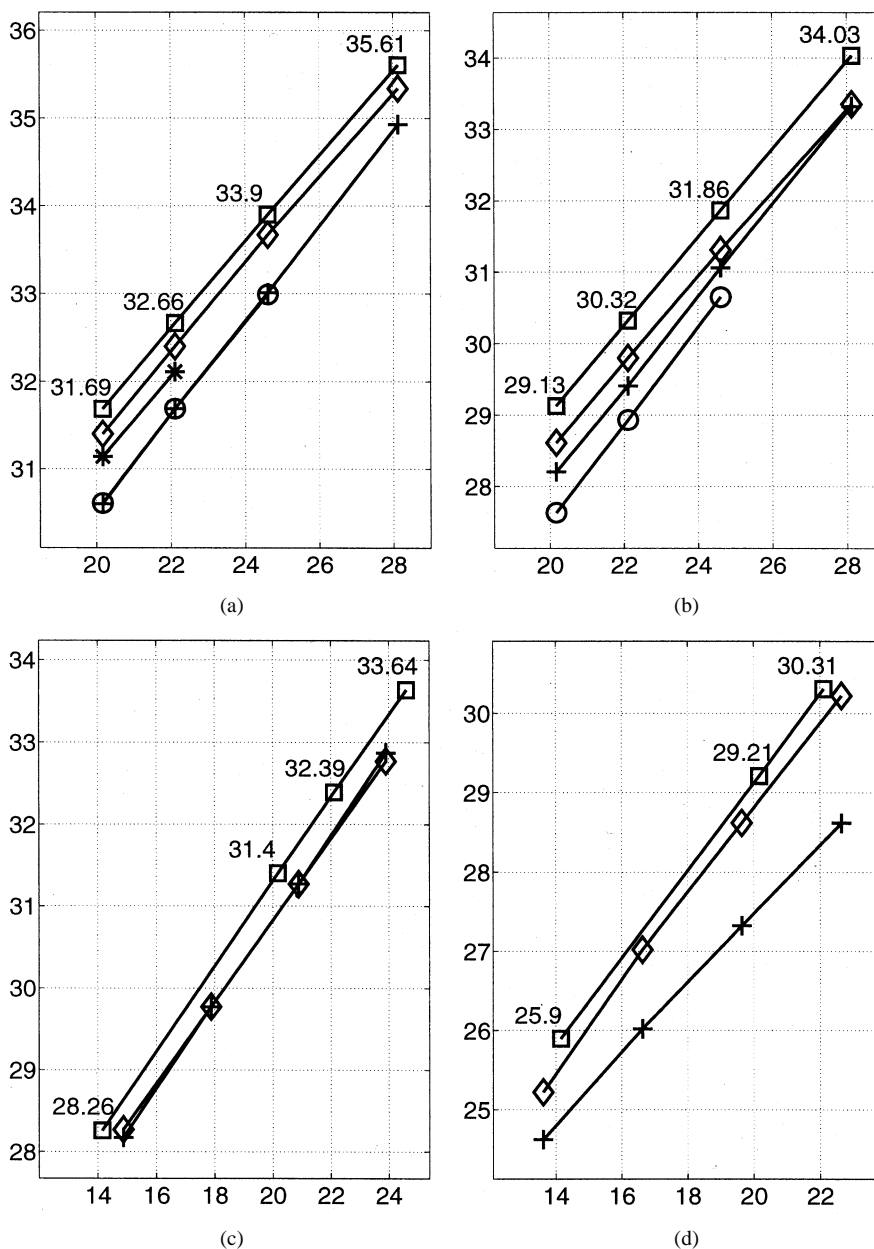


Fig. 5. Comparison of denoising performance of several recently published methods. Curves depict output PSNR as a function of input PSNR. Square symbols indicate our results, taken from Table I. (a,b) circles [32]; crosses [35]; asterisk [52]³; (c,d) crosses [31]; diamonds [51].

undecimated separable wavelet transform. The decrease is substantial, however, in the case of the critically sampled representation. From the comparison of the outcomes of both sets of experiments, one may conclude that both our representation and estimation strategy contribute significantly to the performance advantage shown in Fig. 3(b).

C. Comparison to State-of-the-Art Methods

Finally, we have compared our method to some of the best available published results, and these are shown in Fig. 5. Since there are many different versions of the test images available on the Internet, whenever it was possible we have verified directly with the authors that we are using the same images ([35], [50]–[52]), or have used other authors’ data included in previous comparisons from those authors ([31], [32]) (see Appendix B

for more details about the origin of the images). Fig. 6 provides a visual comparison of an example image (*Barbara*) denoised using the algorithm of Li *et al.* [35], which is based on a variance-adaptive model in an overcomplete separable wavelet representation. Note that the noisy images were created using different samples of noise, and thus the artifacts in the two images appear at different locations. Our method is seen to provide fewer artifacts as well as better preservation of edges and other details. The separation of diagonal orientations in the steerable pyramid allows more selective removal of the noise in diagonally oriented image regions (see parallel diagonal lines on the left side of the face).

³These two plotted PSNR values have been obtained by Starck using the standard Lena image provided by us, which differs from the version used in [52].



Fig. 6. Comparison of denoising results on *Barbara* image (cropped to 150×150 for visibility of the artifacts). From left to right and top to bottom: Original image; Noisy image ($\sigma = 25$, $PSNR = 20.2$ dB); Results of Li *et al.* [35] ($PSNR = 28.2$ dB); Our method ($PSNR = 29.1$ dB).

VI. CONCLUSIONS

We have presented a denoising method based on a local Gaussian scale mixture model in an overcomplete oriented pyramid representation. Our statistical model differs from previous models in a number of important ways. First, many previous models have been based on either separable orthogonal wavelets, or redundant versions of such wavelets. In contrast, our model is based on an overcomplete tight frame that is free from aliasing, and that includes basis functions that are selective for oblique orientations. The increased redundancy of the representation and the higher ability to discriminate orientations results in improved performance. Second, our model explicitly incorporates the covariance between neighboring coefficients (for both signal and noise), as opposed to considering only marginal responses or local variance. Thus, the model captures correlations induced by the overcomplete representation as well as correlations inherent in the underlying image, and it can handle Gaussian noise of arbitrary power spectral density. Third, we have included a neighbor from the same orientation and spatial location at a coarser scale (*a parent*), as opposed to considering only spatial neighbors within each subband. This modeling choice is consistent with the empirical findings of strong statistical dependence across scale in natural images, e.g., [4], [46]. Note, however, that the inclusion of the parent results in only a modest increase in performance compared to the other elements shown in Table II. We believe the impact of including a parent is limited by the simplicity of our model,

which only characterizes the correlation of the coefficients and the correlation of their amplitudes (see below).

In addition to these modeling differences, there are also differences between our denoising method and previous methods based on continuous hidden-variable models [3], [28], [32], [34], [36]. First, we compute the full optimal local Bayesian least squares solution, as opposed to first estimating the local variance, and then using this to estimate the coefficient. We have shown empirically that this approach yields an important improvement in the results. Also, we use the vectorial form of the LLS solution (9), so taking full advantage of the information provided by the covariance modeling of signal and noise. These enhancements, together with a convenient choice for the prior of the hidden multiplier (a noninformative prior, independent of the observed signal), result in a substantial improvement in the quality of the denoised images, while keeping the computational cost reasonably low.

We are currently working on several extensions of the estimator presented here. First, we have begun developing a variant of this method to denoise color images taken with a commercial digital camera [53]. We find that the sensor noise of such cameras has two important features that must be characterized through calibration measurements: spatial and cross-channel correlation, and signal-dependence. We are also extending the denoising solution to address the complete image restoration problem, by incorporating a model of image blur [54]. Finally, we are developing an ML estimator for the noise variance, when the normalized power spectral density

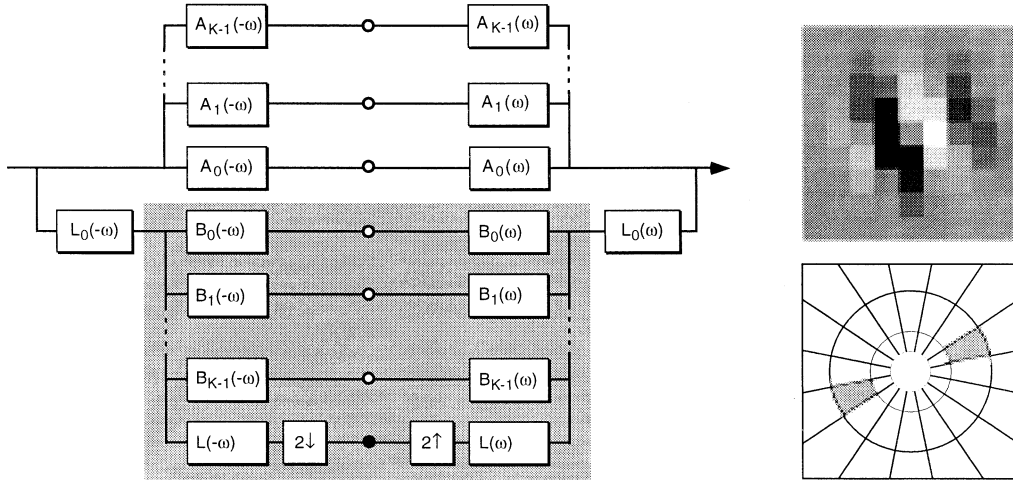


Fig. 7. (a) System diagram for the extended version of the steerable pyramid used in this paper [38]. The input image is first split into a lowpass band and a set of highpass oriented bands. The lowpass band is then split into a lower-frequency band and a set of oriented subbands. The pyramid recursion consists of inserting the diagram contents of the shaded region at the lowpass branch (solid circle). (b) Basis function corresponding to an example oriented subband, and idealized depiction of the frequency domain partition ($K = 8$, $J = 2$), with gray region corresponding to this basis function.

of the noise is assumed known. Preliminary results of these extensions appear very promising.

We believe that the current image model can be improved in a number of ways. It would be desirable to develop a method for efficiently estimating a prior for the multiplier by maximizing the *joint* likelihood of the observed subbands, as opposed to the somewhat heuristic choice of noninformative prior (corrected at the origin) we have presented. Similarly, it would also be desirable to minimize the expected quadratic error in the image domain, instead of doing it for the subband coefficients. In addition, it is worth exploring the transformation of the local GSM model into an explicit Markov model with overlapping neighborhoods, as opposed to the nonoverlapping tree-structured models previously developed [3], [25]. This conceptual simplification would facilitate other applications requiring a conditional local density model (e.g., synthesis or coding). Finally, from a longer-term perspective, major improvements are likely to come from statistical models that capture important structural properties of local image features, by including additional dependencies such as phase congruency between the coefficients of complex multiscale oriented transforms, e.g., [55], [56].

APPENDIX A STEERABLE PYRAMID

We use a transform known as a *steerable pyramid* [38], [39] to decompose images into frequency subbands. The transform is implemented in the Fourier domain, allowing exact reconstruction of the image from the subbands, as well as a flexible choice of the number of orientations (K) and scales (J). A software implementation (in Matlab) is available at <http://www.cns.nyu.edu/~lcv/software.html>. As with conventional orthogonal wavelet decompositions, the pyramid is implemented by recursively splitting an image into a set of oriented subbands, and a lowpass residual band which is subsampled by a factor of two along both axes. Unlike conventional

orthogonal wavelet decompositions, the oriented bands are not subsampled, and the subsampling of the lowpass band does not produce aliasing artifacts, as the lowpass filter is designed to obey the Nyquist sampling criterion. When performing convolutions, the boundaries are handled by mirror extension (reflection) of the image, thereby maintaining continuity. Since it is a tight frame, the transformation may be inverted by convolving each subband with its associated complex-conjugated filter and adding the results. The redundancy factor of this overcomplete representation is (for $J \rightarrow \infty$) $(7/3)K$.

The system diagram for the transform is shown in Fig. 7(a). The filters are polar-separable in the Fourier domain, where they may be written as:

$$L(r, \theta) = \begin{cases} \cos\left(\frac{\pi}{2} \log_2\left(\frac{4r}{\pi}\right)\right), & \frac{\pi}{4} < r < \frac{\pi}{2} \\ 1, & r \leq \frac{\pi}{4} \\ 0, & r \geq \frac{\pi}{2} \end{cases}$$

$$B_k(r, \theta) = H(r)G_k(\theta), \quad k \in [0, K-1]$$

where r, θ are polar frequency coordinates, and

$$H(r) = \begin{cases} \cos\left(\frac{\pi}{2} \log_2\left(\frac{2r}{\pi}\right)\right), & \frac{\pi}{4} < r < \frac{\pi}{2} \\ 1, & r \geq \frac{\pi}{4} \\ 0, & r \leq \frac{\pi}{4} \end{cases}$$

$$G_k(\theta) = \frac{(K-1)!}{\sqrt{K} [2(K-1)]!} \left[2 \cos\left(\theta - \frac{\pi k}{K}\right) \right]^{K-1}.$$

The recursive procedure is initialized by splitting the input image into lowpass and oriented highpass portions, using the following filters:

$$L_0(r, \theta) = L\left(\frac{r}{2}, \theta\right), \quad A_k(r, \theta) = H\left(\frac{r}{2}\right) G_k(\theta).$$

Fig. 7(b) also shows the impulse response of an example band-oriented filter (for $K = 8$), at the highest resolution level, together with its (purely imaginary) Fourier transform.

APPENDIX B
ORIGIN OF THE TEST IMAGES

All 8-bit grayscale test images used in obtaining our results are available on the Internet from <http://decsai.ugr.es/~javier/denoise>. Five of the images, commonly known as *Lena*, *Barbara*, *Boats*, *House* and *Peppers*, are widely used in the image processing literature. Unfortunately, most test images are available in more than one version, with differences between them due to cropping, scanning, resizing, compression or conversion from color to gray-level. In the versions used in this paper the first three are 512×512 and the last two are 256×256 . We also included a 512×512 image of a fingerprint, which unlike the other images, is a homogeneous texture.

Among the several versions of 512×512 -bit gray-level *Lena*, we chose the one that seems the most standard, from <http://www.ece.rice.edu/~wakin/images/Lena512.bmp>. For the comparison of Fig. 5(a), Starck generously offered to run his algorithm on our test image, and Li [35] and Sendur [50] kindly confirmed they were using the same version of the image. The *Barbara* image was obtained from Schmid's standard test images database at <http://ju.sourceforge.net/testimages/index.html>. This version had been previously used in [35], where, in turn, there is a comparison to [32]. It has also been used in [50]. The *Boats* image was taken from University of Southern California SIPI image database at <http://sipi.usc.edu/services/database/database.cgi>. This same version has been used in [50]. The *House* and *Peppers* images were kindly provided by Pižurica, for proper comparison to her results reported in [51].

REFERENCES

- [1] D. Andrews and C. Mallows, "Scale mixtures of normal distributions," *J. R. Statist. Soc.*, vol. 36, p. 99, 1974.
- [2] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," in *Adv. Neural Information Processing Systems*, S. A. Solla, T. K. Leen, and K. R. Müller, Eds. Cambridge, MA: MIT Press, 2000, vol. 12, pp. 855–861.
- [3] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in modeling and analyzing natural imagery," *Appl. Comput. Harmon. Anal.*, vol. 11, no. 1, pp. 89–123, July 2001.
- [4] D. L. Ruderman, "The statistics of natural images," *Network: Comput. Neural Syst.*, vol. 5, pp. 517–548, 1996.
- [5] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Amer. A*, vol. 4, no. 12, pp. 2379–2394, 1987.
- [6] J. G. Daugman, "Entropy reduction and decorrelation in visual coding by oriented neural receptive fields," *IEEE Trans. Biomed. Eng.*, vol. 36, no. 1, pp. 107–114, 1989.
- [7] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Pattern Anal. Machine Intell.*, vol. 11, pp. 674–693, July 1989.
- [8] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.
- [9] A. J. Bell and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vis. Res.*, vol. 37, no. 23, pp. 3327–3338, 1997.
- [10] J. P. Rossi, "Digital techniques for reducing television noise," *JSMPT*, vol. 87, pp. 134–140, 1978.
- [11] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.
- [12] D. Leporini and J. C. Pesquet, "Multiscale regularization in Besov spaces," in *Proc. 31st Asilomar Conf. on Signals, Systems and Computers*, Pacific Grove, CA, Nov. 1998.
- [13] E. P. Simoncelli and E. H. Adelson, "Noise removal via Bayesian wavelet coring," in *Proc. 3rd Int. Conf. on Image Processing*, vol. I, Lausanne, Switzerland, Sept. 1996, pp. 379–382.
- [14] H. A. Chipman, E. D. Kolaczyk, and R. M. McCulloch, "Adaptive Bayesian wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 92, no. 440, pp. 1413–1421, 1997.
- [15] F. Abramovich, T. Sapatinas, and B. W. Silverman, "Wavelet thresholding via a Bayesian approach," *J. R. Statist. Soc. B*, vol. 60, pp. 725–749, 1998.
- [16] E. P. Simoncelli, "Bayesian denoising of visual images in the wavelet domain," in *Bayesian Inference in Wavelet Based Models*, P. Müller and B. Vidakovic, Eds. New York: Springer-Verlag, 1999, vol. 141, ch. 18, pp. 291–308.
- [17] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using a generalized Gaussian and complexity priors," *IEEE Trans. Inform. Theory*, vol. 45, pp. 909–919, 1999.
- [18] A. Hyvarinen, "Sparse code shrinkage: Denoising of non-Gaussian data by maximum likelihood estimation," *Neural Comput.*, vol. 11, no. 7, pp. 1739–1768, 1999.
- [19] J. Starck, E. J. Candes, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Processing*, vol. 11, pp. 670–684, June 2002.
- [20] B. Wegmann and C. Zetsche, "Statistical dependence between orientation filter outputs used in an human vision based image code," in *Proc. Visual Comm. Image Processing*, vol. 1360, Lausanne, Switzerland, 1990, pp. 909–922.
- [21] E. P. Simoncelli, "Statistical models for images: Compression, restoration and synthesis," presented at Proc. 31st Asilomar Conf. on Signals, Systems and Computers. [Online]. Available: <http://www.cns.nyu.edu/~eero/publications.html>
- [22] R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Trans. Image Processing*, vol. 8, pp. 1688–1701, Dec. 1999.
- [23] H. Brehm and W. Stammers, "Description and generation of spherically invariant speech-model signals," *Signal Process.*, vol. 12, pp. 119–141, 1987.
- [24] T. Bollersley, K. Engle, and D. Nelson, "ARCH models," in *Handbook of Econometrics V*, B. Engle and D. McFadden, Eds., 1994.
- [25] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 46, pp. 886–902, Apr. 1998.
- [26] J. Romberg, H. Choi, and R. Baraniuk, "Bayesian tree-structured image modeling using Wavelet-domain hidden Markov models," *IEEE Trans. Image Processing*, vol. 10, July 2001.
- [27] S. M. LoPresto, K. Ramchandran, and M. T. Orchard, "Wavelet image coding based on a new generalized Gaussian mixture model," in *Proc. Data Compression Conf.*, Snowbird, UT, Mar. 1997.
- [28] M. K. Mihçak, I. Kozintsev, K. Ramchandran, and P. Moulin, "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 6, pp. 300–303, Dec. 1999.
- [29] J. S. Lee, "Digital image enhancement and noise filtering by use of local statistics," *IEEE Pattern Anal. Machine Intell.*, vol. PAMI-2, pp. 165–168, Mar. 1980.
- [30] H. Robbins, "The empirical Bayes approach to statistical decision problems," *Ann. Math. Statist.*, vol. 35, pp. 1–20, 1964.
- [31] M. Malfait and D. Roose, "Wavelet-based image denoising using a Markov random field a priori model," *IEEE Trans. Image Processing*, vol. 6, pp. 549–565, Apr. 1997.
- [32] S. G. Chang, B. Yu, and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," in *Proc. 5th IEEE Int. Conf. Image Processing*, Chicago, IL, Oct. 1998.
- [33] F. Abramovich, T. Besbeas, and T. Sapatinas, "Empirical Bayes approach to block wavelet function estimation," *Comput. Statist. Data Anal.*, vol. 39, pp. 435–451, 2002.
- [34] V. Strela, J. Portilla, and E. Simoncelli, "Image denoising using a local Gaussian scale mixture model in the wavelet domain," *Proc. SPIE Wavelet Applications in Signal and Image Processing VIII*, vol. 4119, pp. 363–371, Dec. 2000.
- [35] X. Li and M. T. Orchard, "Spatially adaptive image denoising under overcomplete expansion," in *Proc. IEEE Int. Conf. Image Processing*, Vancouver, BC, Canada, Sept. 2000.
- [36] J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, "Adaptive Wiener denoising using a Gaussian scale mixture model in the wavelet domain," in *Proc. 8th IEEE Int. Conf. Image Processing*, Thessaloniki, Greece, Oct. 7–10, 2001, pp. 37–40.

- [37] R. R. Coifman and D. L. Donoho, "Translation-invariant de-noising," in *Wavelets and Statistics*, A. Antoniadis and G. Oppenheim, Eds. San Diego, CA: Springer-Verlag, 1995, Lecture notes.
- [38] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans. Inform. Theory*, vol. 38, pp. 587–607, Mar. 1992.
- [39] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Pattern Anal. Machine Intell.*, vol. 13, no. 9, pp. 891–906, 1991.
- [40] A. B. Watson, "The cortex transform: Rapid computation of simulated neural images," *Comput. Vis. Graphics Image Process.*, vol. 39, pp. 311–327, 1987.
- [41] E. J. Candes and D. L. Donoho, "Ridgelets: A key to higher-dimensional intermittency?," *Phil. Trans. R. Soc. Lond A*, vol. 357, pp. 2495–2509, 1999.
- [42] N. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Appl. Comput. Harmon. Anal.*, vol. 10, no. 3, pp. 234–253, May 2001.
- [43] V. Strela, "Denoising via block Wiener filtering in wavelet domain," in *Proc. 3rd Eur. Congr. Math.*, Barcelona, Spain, July 2000.
- [44] G. E. P. Box and C. Tiao, *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley, 1992.
- [45] M. Figueiredo and R. Nowak, "Wavelet-based image estimation: An empirical Bayes approach using Jeffrey's noninformative prior," *IEEE Trans. Image Processing*, vol. 10, pp. 1322–1331, Sept. 2001.
- [46] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.
- [47] D. L. Ruderman, "Origins of scaling in natural images," *Vis. Res.*, vol. 37, pp. 3385–3398, 1997.
- [48] A. Taberero, J. Portilla, and R. Navarro, "Duality of log-polar image representations in the space and the spatial-frequency domains," *IEEE Trans. Signal Processing*, vol. 47, pp. 2469–2479, Sept. 1999.
- [49] L. Sendur and I. W. Selesnick, "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency," *IEEE Trans. Signal Processing*, vol. 50, pp. 2744–2756, Nov. 2002.
- [50] —, "Bivariate shrinkage with local variance estimation," *IEEE Signal Processing Lett.*, vol. 9, pp. 438–441, Dec. 2002.
- [51] A. Pižurica, W. Philips, I. Lemahieu, and M. Acheroy, "A joint inter- and intrascale statistical model for Bayesian wavelet based image denoising," *IEEE Trans. Image Processing*, vol. 11, pp. 545–557, May 2002.
- [52] J. L. Starck, D. L. Donoho, and E. Candes, "Very high quality image restoration," *Proc. SPIE*, vol. 4478, pp. 9–19, August 2001.
- [53] J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, "Image Denoising Using Gaussian Scale Mixtures in the Wavelet Domain," Courant Inst. of Math. Sci., New York University, Tech. Rep. TR2002-831, 2002.
- [54] J. Portilla and E. P. Simoncelli, "Image restoration using Gaussian scale mixtures in the wavelet domain," in *Proc. IEEE Int. Conf. on Image Proc.*, Barcelona, Spain, Sept. 2003.
- [55] P. D. Kovesi, "Image features from phase congruency," *J. Comput. Vis. Res.*, vol. 1, no. 3, Summer 1999.
- [56] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int. J. Comput. Vis.*, vol. 40, no. 1, pp. 49–71, 2000.
- [57] A. Chambolle, R. A. DeVore, N. Lee, and B. J. Lucier, "Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. Image Processing*, vol. 7, pp. 319–335, Mar. 1998.
- [58] B. Vidakovic, "Nonlinear wavelet shrinkage with Bayes rules and Bayes factors," *J. Amer. Statist. Assoc.*, vol. 93, pp. 173–179, 1998.



Javier Portilla received the M.S. degree in 1994 and the Ph.D. degree in 1999, both in electrical engineering, from the Universidad Politécnica de Madrid.

From 1995 to 1999, he was a research assistant at the Instituto de Óptica, Consejo Superior de Investigaciones Científicas, Madrid. From 1999 to 2001, he was a Research Associate in E. P. Simoncelli's laboratory, at the Center for Neural Science, New York University. Currently he is an Associate Investigator within the Visual Information Processing Group, at the Computer Science and Artificial Intelligence Department of the Universidad de Granada. His research is focused on visual-statistical representation models for natural images and textures, and their application to image processing and synthesis.



Vasily Strela received the Ph.D. from the Massachusetts Institute of Technology.

He held visiting positions at the University of South Carolina, Imperial College, Dartmouth College, New York University, and an assistant professorship at Drexel University. Currently he is working in industry. His research interests include wavelet and multiwavelet theory, signal processing, and financial mathematics.



Martin J. Wainwright received the Ph.D. degree in electrical engineering and computer science (EECS) from the Massachusetts Institute of Technology (MIT), Cambridge, in January 2002.

He is currently a Postdoctoral Research Associate in EECS at the University of California, Berkeley. His interests include statistical signal and image processing, variational methods and convex optimization, machine learning, and information theory.

Dr. Wainwright received the George M. Sprowls Award from the MIT EECS Department for his doctoral dissertation.

tor dissertation.



Eero P. Simoncelli received the B.S. degree in physics in 1984 from Harvard University, Cambridge, MA. He studied applied mathematics at Cambridge University for a year and a half, and received the M.S. degree in 1988 and the Ph.D. degree in 1993, both in electrical engineering from the Massachusetts Institute of Technology.

He was an Assistant Professor with the Computer and Information Science Department at the University of Pennsylvania, Philadelphia, from 1993 until 1996. He moved to New York University in September of 1996, where he is currently an Associate Professor in neural science and mathematics. In August 2000, he became an Associate Investigator of the Howard Hughes Medical Institute, under their new program in computational biology. His research interests span a wide range of issues in the representation and analysis of visual images, in both machine and biological vision systems.