Bivariate Shrinkage Functions for Wavelet-Based Denoising Exploiting Interscale Dependency

Levent Şendur, *Student Member, IEEE*, and Ivan W. Selesnick, *Member, IEEE*

***Abstract*—Most simple nonlinear thresholding rules for wavelet-based denoising assume that the wavelet coefficients are independent. However, wavelet coefficients of natural images have significant dependencies. In this paper, we will only consider the dependencies between the coefficients and their parents in detail. For this purpose, new non-Gaussian bivariate distributions are proposed, and corresponding nonlinear threshold functions (shrinkage functions) are derived from the models using Bayesian estimation theory. The new shrinkage functions do not assume the independence of wavelet coefficients. We will show three image denoising examples in order to show the performance of these new bivariate shrinkage rules. In the second example, a simple subband-dependent data-driven image denoising system is described and compared with effective data-driven techniques in the literature, namely VisuShrink, SureShrink, BayesShrink, and hidden Markov models. In the third example, the same idea is applied to the dual-tree complex wavelet coefficients.**

***Index Terms*—Bivariate shrinkage, image denoising, statistical modeling, wavelet transforms.**

I. Introduction

MULTISCALE decompositions have shown significant advantages in the representation of signals, and they are used extensively in image compression [9], [32], segmentation [6], [25], and denoising [10], [27], [29], [30], for example. In this paper, we will mostly deal with the modeling of the wavelet transform coefficients of natural images and its application to the image denoising problem. The denoising of a natural image corrupted by Gaussian noise is a classic problem in signal processing. The wavelet transform has become an important tool for this problem due to its energy compaction property. Crudely, it states that the wavelet transform yields a large number of small coefficients and a small number of large coefficients.

Simple denoising algorithms that use the wavelet transform consist of three steps.

```
1) Calculate the wavelet transform of the
noisy signal.
2) Modify the noisy wavelet coefficients
according to some rule.
3) Compute the inverse transform using the
modified coefficients.
```

Manuscript received January 24, 2002; revised June 20 2002. This work was supported by the National Science Foundation under CAREER Grant CCR-9875452. The associate editor coordinating the review of this paper and approving it for publication was Dr. Xiang-Gen Xia.

The authors are with Electrical and Computer Engineering, Polytechnic University, Brooklyn, NY 11201 USA (e-mail: levent@taco.poly.edu; selesi@taco.poly.edu).

Digital Object Identifier 10.1109/TSP.2002.804091.

One of the most well-known rules for the second step is soft thresholding analyzed by Donoho [13]. Due to its effectiveness and simplicity, it is frequently used in the literature. The main idea is to subtract the threshold value T from all coefficients larger than T and to set all other coefficients to zero. Alternative approaches can be found in, for example, [1], [2], [5], [7], [10], [11], [14], [15], [17], [18], [21], [22], [27], [29], [30], [35], [36], [38], and [39]. Generally, these methods use a threshold value that must be estimated correctly in order to obtain good performance. VisuShrink [14] uses one of the well-known thresholding rules: the universal threshold. In addition, subband adaptive systems have superior performance, such as SureShrink [15], which is a data-driven system. Recently, BayesShrink [4], which is also a data-driven subband adaptive technique, is proposed and outperfoms VisuShrink and SureShrink.

Recently, some research has addressed the development of statistical models for natural images and their transform coefficients [16], [19], [20], [33], [34], [36]. Hence, statistical approaches have emerged as a new tool for wavelet-based denoising. The basic idea is to model wavelet transform coefficients with prior probability distributions. Then, the problem can be expressed as the estimation of clean coefficients using this *a priori* information with Bayesian estimation techniques, such as the MAP estimator.

If the MAP estimator is used for this problem, the solution requires *a priori* knowledge about the distribution of wavelet coefficients. Therefore, two problems arise: 1) What kind of distributions represent the wavelet coefficients and 2) what is the corresponding estimator (shrinkage function)?

Statistical models pretend wavelet coefficients are random variables described by some probability distribution function. For the first problem, these models mostly assume that the coefficients are independent and try to characterize them by using Gaussian, Laplacian, generalized Gaussian, or other distributions. For example, the classical soft threshold shrinkage function can be obtained by a Laplacian assumption. Bayesian methods for image denoising using other distributions have also been proposed [17], [18], [21], [22], [37], [39]. However, these simple distributions are weak models for wavelet coefficients of natural images because they ignore the dependencies between coefficients. It is well known that wavelet coefficients are statistically dependent [35] due to two properties of the wavelet transform: 1) If a wavelet coefficient is large/small, the adjacent coefficients are likely to be large/small, and 2) large/small coefficients tend to propagate across the scales.

Algorithms that exploit the dependency between coefficients can give better results compared with the ones derived using an independence assumption [5], [7], [11], [12], [26], [30], [35], [36], [38], [41], [42]. For example, in [11], a new framework

to capture the statistical dependencies by using wavelet-domain hidden Markov models (HMT) is developed. In [41] and [42], improved local contextual hidden Markov models are introduced. In [35] and [36], the wavelet coefficients are assumed to be Gaussian given the neighbor coefficients, and a joint shrinkage function that uses neighbor wavelet coefficients is proposed. In [30], [38], and [40], the class of Gaussian scale mixtures for modeling natural images is proposed and applied to the image-denoising problem. A local adaptive window-based image denoising algorithm using ML and MAP estimates, which is a powerful low-complexity algorithm exploiting the intrascale dependencies of wavelet coefficients [29], is proposed. Interscale dependencies are considered as an extension of this algorithm in [3]. In addition, [28] presents an information-theoretic analysis of statistical dependencies between wavelet coefficients. In [12], a Bayesian approach where the underlying signal in possibly non-Gaussian noise is modeled using Besov norm priors is proposed. In this paper, we use new bivariate probability distribution functions (pdfs) and derive corresponding bivariate shrinkage functions using Bayesian estimation theory, specifically the MAP estimator. In this paper, the aim is similar to [12]; however, a general Besov norm is not used here as it is in [12], but an explicit formula for a shrinkage function is obtained. An explicit multivariate shrinkage function for wavelet denoising is also presented in [44]. The shrinkage function in [44], which was derived using a different method, is similar to Model 1, which is one of the shinkage methods that will be presented in this paper.

The organization of this paper is as follows. In Section II, the basic idea of Bayesian denoising will be briefly described. Section II-A describes the marginal models and how the soft thresholding operator can be obtained using the Laplace pdf. Then, Section II-B describes the bivariate models, and new non-Gaussian bivariate pdfs are proposed to model the joint statistics of wavelet coefficients. These models try to capture the dependencies between a coefficient and its parent in detail. Four new pdfs will be proposed, starting from the simplest model, and more complicated models are proposed in order to characterize a larger family of pdfs. In addition, for each model, the corresponding MAP estimator will be obtained. In Section III, three image denoising examples based on these new models will be illustrated and compared with other algorithms. In the first example, the new bivariate shrinkage function developed using Model 1 and the classical soft thresholding will be compared by optimizing the threshold value. In the second example, a simple subband adaptive data-driven image denoising algorithm that uses Model 3 will be described, and the estimation of the parameters, which is necessary for the model, will be explained. The results will be compared with the VisuShrink, SureShrink, BayesShrink, and HMT algorithms. In the third example, the performance of a subband-dependent system will be demonstrated on the dual-tree complex wavelet transform.

II. Bayesian Denoising

In this section, the denoising of an image corrupted by white Gaussian noise will be considered, i.e.,

$$g = x + n \tag{1}$$

where n is independent Gaussian noise. We observe g (a noisy signal) and wish to estimate the desired signal x as accurately as possible according to some criteria. In the wavelet domain, if we use an orthogonal wavelet transform, the problem can be formulated as

$$y = w + n \tag{2}$$

where

- y noisy wavelet coefficient;
- w true coefficient;
- n noise, which is independent Gaussian.

This is a classical problem in estimation theory. Our aim is to estimate w from the noisy observation y. The maximum *a posteriori* (MAP) estimator will be used for this purpose. Marginal and bivariate models will be discussed for this problem in Sections II-A and B, and new MAP estimators are derived.

A. Marginal Model

The classical MAP estimator for (2) is

$$\hat{w}(y) = \arg\max_w p_{w|y}(w|y). \tag{3}$$

Using Bayes rule, one gets

$$\begin{aligned}\hat{w}(y) &= \arg\max_w \left[p_{y|w}(y|w) \cdot p_w(w)\right] \\ &= \arg\max_w \left[p_n(y-w) \cdot p_w(w)\right].\end{aligned} \tag{4}$$

Therefore, these equations allow us to write this estimation in terms of the pdf of the noise (p_n) and the pdf of the signal coefficient (p_w). From the assumption on the noise, p_n is zero mean Gaussian with variance σ_n, i.e.,

$$p_n(n) = \frac{1}{\sigma_n\sqrt{2\pi}} \cdot \exp\left(-\frac{n^2}{2\sigma_n^2}\right). \tag{5}$$

It has been observed that wavelet coefficients of natural images have highly non-Gaussian statistics [16], [33]. The pdf for wavelet coefficients is often modeled as a generalized (heavy-tailed) Gaussian [35]

$$p_w(w) = K(s,p) \cdot \exp\left(-\left|\frac{w}{s}\right|^p\right) \tag{6}$$

where s, p are the parameters for this model, and $K(s,p)$ is the parameter-dependent normalization constant. Other pdf models have also been proposed [17], [18], [21], [22], [39]. In practice, generally, two problems arise with the Bayesian approach when an accurate but complicated pdf $p_w(w)$ is used: 1) It can be difficult to estimate the parameters of p_w for a specific image, especially from noisy data, and 2) the estimators for these models may not have simple closed form solution and can be difficult to obtain. The solution for these problems usually requires numerical techniques.

Fig. 1(a) illustrates the histogram of the wavelet coefficients computed from several natural images. A Laplacian density is fitted to this empirical histogram. The same data is plotted in the log domain in Fig. 1(b) in order to show the failure of Laplacian assumption along the tail. Even though it is not the most accurate, the Laplacian is an especially convenient assumption for

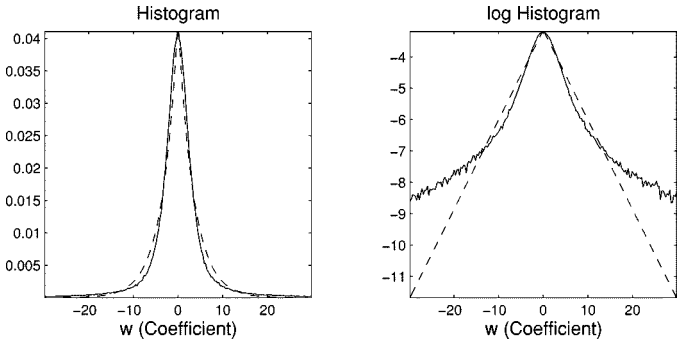

Fig. 1. (a) Empirical histogram computed from several natural images (solid line). A Laplacian pdf is fitted to the empirical histogram (dashed line). (b) Same data is illustrated in log domain in order to emphasize the tail difference.

$p_w(w)$ because the MAP estimator is simple to compute and is, in fact, given by the soft threshold rule.

Let us continue developing the MAP estimator and show it for Gaussian and Laplacian cases. Equation (4) is also equivalent to

$$\hat{w}(y) = \arg\max_w \left[\log(p_n(y-w)) + \log(p_w(w))\right]. \quad (7)$$

As in [21], let us define $f(w) = \log(p_w(w))$. By using (5), (7) becomes

$$\hat{w}(y) = \arg\max_w \left[-\frac{(y-w)^2}{2\sigma_n^2} + f(w)\right]. \quad (8)$$

This is equivalent to solving the following equation for $\hat{w}$ if $p_w(w)$ is assumed to be strictly convex and differentiable.

$$\frac{y-\hat{w}}{\sigma_n^2} + f'(\hat{w}) = 0. \quad (9)$$

If $p_w(w)$ is assumed to be a zero mean Gaussian density with variance σ^2, then $f(w) = -\log(\sqrt{2\pi}\sigma) - w^2/2\sigma^2$, and the estimator can be written as

$$\hat{w}(y) = \frac{\sigma^2}{\sigma^2+\sigma_n^2} \cdot y. \quad (10)$$

If it is Laplacian

$$p_w(w) = \frac{1}{\sqrt{2}\sigma} \exp\left(-\frac{\sqrt{2}|w|}{\sigma}\right) \quad (11)$$

then $f(w) = -\log(\sigma\sqrt{2}) - \sqrt{2}|w|/\sigma$, and the estimator will be

$$\hat{w}(y) = \text{sign}(y)\left(|y| - \frac{\sqrt{2}\sigma_n^2}{\sigma}\right)_+. \quad (12)$$

Here, $(g)_+$ is defined as

$$(g)_+ = \begin{cases} 0, & \text{if } g < 0 \\ g, & \text{otherwise.} \end{cases} \quad (13)$$

Equation (12) is the classical soft shrinkage function. The Laplacian pdf and corresponding shrinkage function are illustrated in Fig. 2. Let us define the soft operator as

$$\text{soft}(g,\tau) = \text{sign}(g) \cdot (|g| - \tau)_+. \quad (14)$$

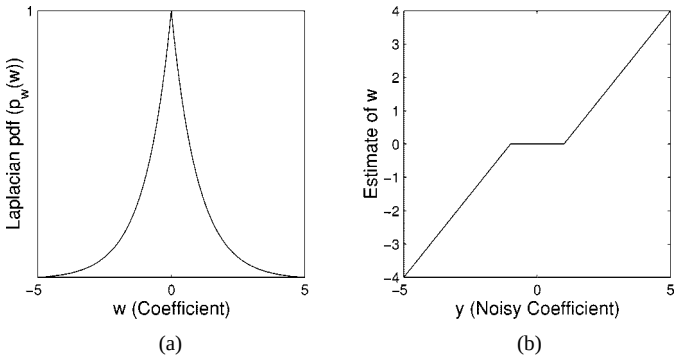

Fig. 2. (a) Laplacian pdf. (b) Corresponding shrinkage function.

This operator will be often used while developing bivariate shrinkage functions. The soft shrinkage function (12) can be written as

$$\hat{w}(y) = \text{soft}\left(y, \frac{\sqrt{2}\sigma_n^2}{\sigma}\right). \quad (15)$$

B. Bivariate Models

Marginal models cannot model the statistical dependencies between wavelet coefficients. However, there are strong dependencies between neighbor coefficients such as between a coefficient, its parent (adjacent coarser scale locations), and their siblings (adjacent spatial locations). In this section, we focus on the dependencies only between a coefficient and its parent in detail. In [35], it is suggested that the pdf of a coefficient, conditioned on neighbor coefficients, is Gaussian, and a linear Bayesian estimator is proposed that requires the estimation of neighbor coefficients. This paper suggests four new jointly non-Gaussian models to characterize the dependency between a coefficient and its parent and derives the corresponding bivariate MAP estimators based on noisy wavelet coefficients in detail.

Here, we modify the Bayesian estimation problem as to take into account the statistical dependency between a coefficient and its parent. Let w_2 represent the parent of w_1. (w_2 is the wavelet coefficient at the same position as w_1, but at the next coarser scale.) Then

$$\begin{aligned} y_1 &= w_1 + n_1 \\ y_2 &= w_2 + n_2 \end{aligned} \quad (16)$$

where y_1 and y_2 are noisy observations of w_1 and w_2, and n_1 and n_2 are noise samples. We can write

$$\mathbf{y} = \mathbf{w} + \mathbf{n} \quad (17)$$

where $\mathbf{w} = (w_1, w_2)$, $\mathbf{y} = (y_1, y_2)$ and $\mathbf{n} = (n_1, n_2)$.

The standard MAP estimator for $\mathbf{w}$ given the corrupted observation $\mathbf{y}$ is

$$\hat{\mathbf{w}}(\mathbf{y}) = \arg\max_{\mathbf{w}} p_{\mathbf{w}|\mathbf{y}}(\mathbf{w}|\mathbf{y}). \quad (18)$$

After some manipulations, this equation can be written as

$$\begin{aligned} \hat{\mathbf{w}}(\mathbf{y}) &= \arg\max_{\mathbf{w}} \left[p_{\mathbf{y}|\mathbf{w}}(\mathbf{y}|\mathbf{w}) \cdot p_{\mathbf{w}}(\mathbf{w})\right] \\ &= \arg\max_{\mathbf{w}} \left[p_{\mathbf{n}}(\mathbf{y}-\mathbf{w}) \cdot p_{\mathbf{w}}(\mathbf{w})\right]. \end{aligned} \quad (19)$$

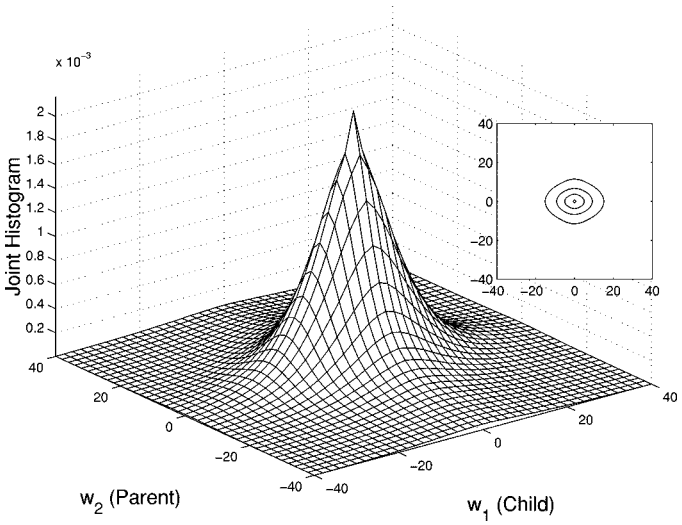

Fig. 3. Empirical joint parent-child histogram of wavelet coefficients (computed from the Corel image database).

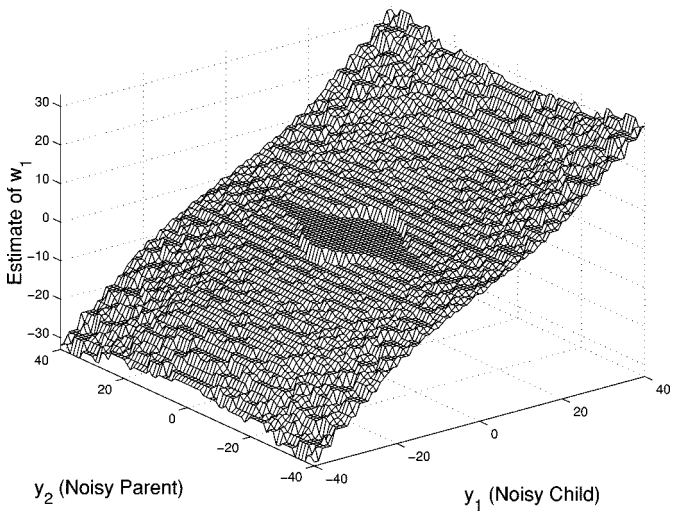

Fig. 4. Joint shrinkage function derived numerically from the empirical joint parent-child histogram illustrated in Fig. 3.

From this equation, the Bayes rule allows us to write this estimation in terms of the probability densities of noise and the prior density of the wavelet coefficients. In order to use this equation to estimate the original signal, we must know both pdfs. We assume the noise is i.i.d. Gaussian, and we write the noise pdf as

$$p_{\mathbf{n}}(\mathbf{n}) = \frac{1}{2\pi\sigma_n^2} \cdot \exp\left(-\frac{n_1^2 + n_2^2}{2\sigma_n^2}\right). \tag{20}$$

The same problem as in marginal case appears. What kind of joint pdf models the wavelet coefficients? The joint empirical coefficient-parent histogram can be used to observe $p_{\mathbf{w}}(\mathbf{w})$. For this purpose, we used 200 512×512 images from the Corel image database in order to stabilize the corresponding statistic. Hence, we got a very smooth histogram. We used Daubechies length-8 filter to compute the wavelet transform. The joint histogram, computed using this set, is illustrated in Fig. 3. Its contour plot is also shown in this figure.

Before proposing models for this empirical histogram, let us investigate what kind of shrinkage function the empirical histogram has. Using empirical data with (19), the shrinkage function can be found numerically. The numerically calculated shrinkage function is illustrated in Fig. 4. As this plot shows, the shrinkage function should depend on both y_1 and y_2.

1) Model 1: It is hard to find a model for the empirical histogram in Fig. 3, but we propose the following pdf:

$$p_{\mathbf{w}}(\mathbf{w}) = \frac{3}{2\pi\sigma^2} \cdot \exp\left(-\frac{\sqrt{3}}{\sigma}\sqrt{w_1^2 + w_2^2}\right). \tag{21}$$

With this pdf, w_1 and w_2 are uncorrelated but not independent. We will call this model Model 1 in order not to confuse this model with the models proposed later in this paper. This is a circularly symmetric pdf and is related to the family of spherically invariant random processes (SIRPs) that are used, for example, in speech processing [31], [43].

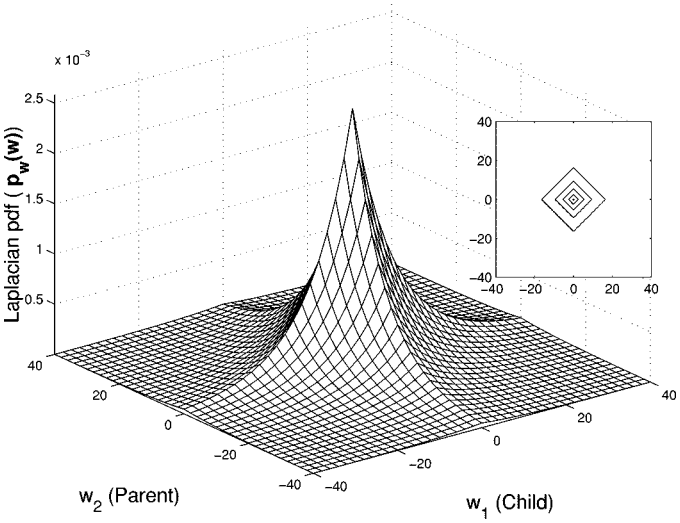

Fig. 5. Independent Laplacian model (22) for joint pdf of parent-child wavelet coefficient pairs.

Before going further with this new model, let us consider the case where w_1 and w_2 are assumed to be independent Laplacian; then, the joint pdf can be written as

$$p_{\mathbf{w}}(\mathbf{w}) = \frac{1}{2\sigma^2} \cdot \exp\left(-\frac{\sqrt{2}}{\sigma}(|w_1| + |w_2|)\right). \tag{22}$$

A plot of this model is illustrated in Fig. 5. If this model is compared with Fig. 3, the difference between them can be easily observed. Let us consider our new model given in (21). The plot of this pdf and its contour plot is illustrated in Fig. 6. As one can easily notice, this model is a much better approximation to the empirical histogram illustrated in Fig. 3.

Let us continue on developing the MAP estimator given in (19), which is equivalent to

$$\hat{\mathbf{w}}(\mathbf{y}) = \arg\max_{\mathbf{w}} \left[\log(p_{\mathbf{n}}(\mathbf{y} - \mathbf{w})) + \log(p_{\mathbf{w}}(\mathbf{w}))\right]. \tag{23}$$

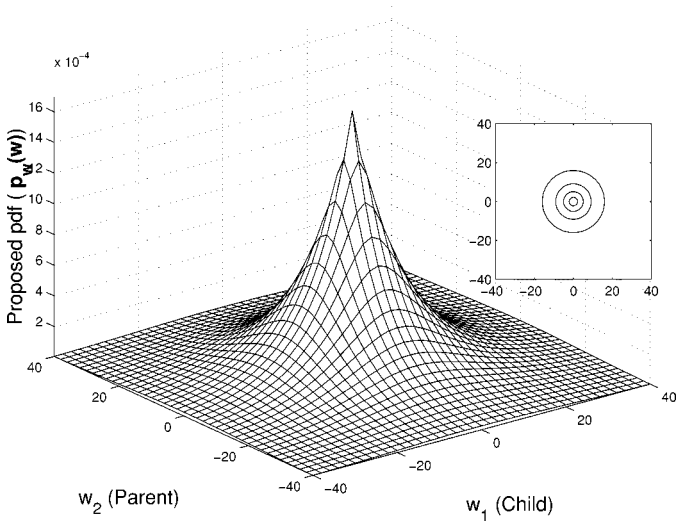

Fig. 6. New bivariate pdf (21) proposed for joint pdf of parent-child wavelet coefficient pairs (Model 1).

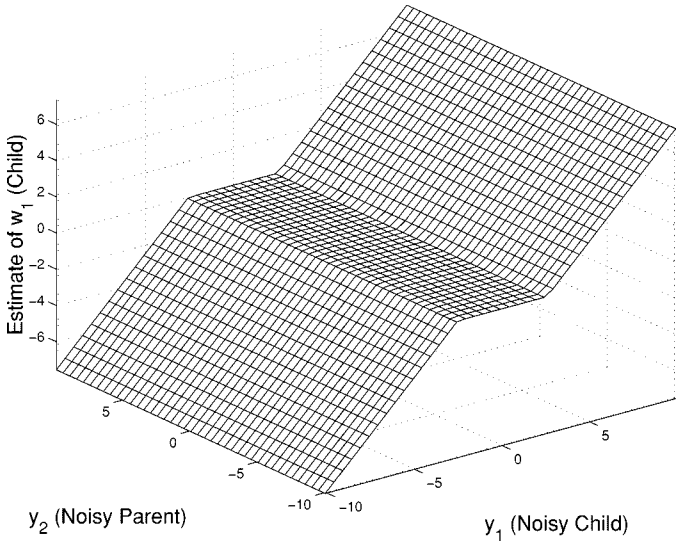

Fig. 7. Joint shrinkage function derived from the Laplacian independent model (Fig. 5).

Let us define $f(\mathbf{w}) = \log(p_{\mathbf{w}}(\mathbf{w}))$. By using (20), (23) becomes

$$\hat{\mathbf{w}}(\mathbf{y}) = \arg\max_{\mathbf{w}} \left[-\frac{(y_1 - w_1)^2}{2\sigma_n^2} - \frac{(y_2 - w_2)^2}{2\sigma_n^2} + f(\mathbf{w}) \right]. \tag{24}$$

This is equivalent to solving the following equations together, if $p_{\mathbf{w}}(\mathbf{w})$ is assumed to be strictly convex and differentible:

$$\frac{y_1 - \hat{w}_1}{\sigma_n^2} + f_1(\hat{\mathbf{w}}) = 0 \tag{25}$$

$$\frac{y_2 - \hat{w}_2}{\sigma_n^2} + f_2(\hat{\mathbf{w}}) = 0 \tag{26}$$

where f_1 and f_2 represent the derivative of $f(w)$ with respect to w_1 and w_2, respectively.

The MAP estimator for the Laplacian independent model in (22) is illustrated in Fig. 7. This rule applies the soft threshold function to y_1 to estimate w_1. It is the usual soft shrinkage function.

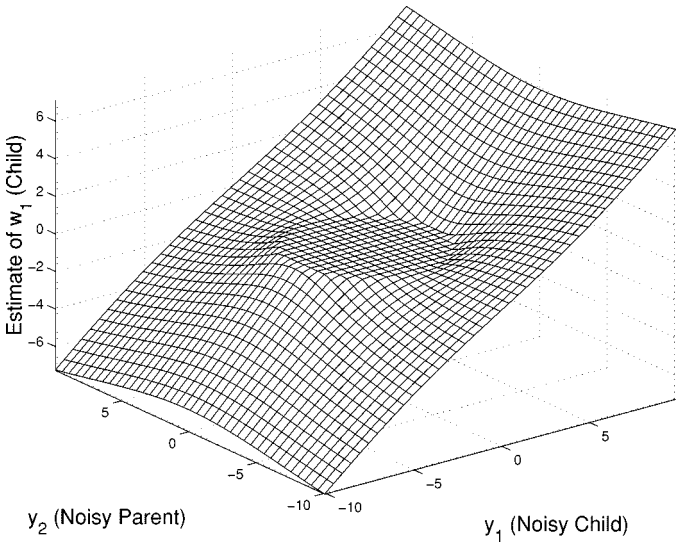

Fig. 8. New bivariate shrinkage function derived from the Model 1 proposed in (21) (Fig. 6).

Let us find the MAP estimator corresponding to our new model given in (21). $f(\mathbf{w})$ can be written as

$$f(\mathbf{w}) = \log\left(\frac{3}{2\pi\sigma^2}\right) - \frac{\sqrt{3}}{\sigma}\sqrt{w_1^2 + w_2^2}. \tag{27}$$

From this

$$f_1(\mathbf{w}) = -\frac{\sqrt{3}w_1}{\sigma\sqrt{w_1^2 + w_2^2}} \tag{28}$$

$$f_2(\mathbf{w}) = -\frac{\sqrt{3}w_2}{\sigma\sqrt{w_1^2 + w_2^2}}. \tag{29}$$

Solving (25) and (26) by using (28) and (29), the MAP estimator (or "the joint shrinkage function") can be written as

$$\hat{w}_1 = \frac{\left(\sqrt{y_1^2 + y_2^2} - \frac{\sqrt{3}\sigma_n^2}{\sigma}\right)_+}{\sqrt{y_1^2 + y_2^2}} \cdot y_1. \tag{30}$$

The derivation can be found in Appendix A. Fig. 8 shows the plot of this bivariate shrinkage function. As this plot illustrates, there is a circular deadzone (the deadzone is the region where the estimated value is zero), i.e.,

$$\text{deadzone} = \left\{ (y_1, y_2) : \sqrt{y_1^2 + y_2^2} \le \frac{\sqrt{3}\sigma_n^2}{\sigma} \right\}.$$

Denoising methods derived using the independence assumption disregard the parent value (y_2) when estimating each coefficient (y_1). For example, in scalar soft thresholding, for all coefficients, the threshold value is fixed and independent from other coefficients—if the coefficient is below the threshold value, we make it zero. However, our results clearly show that the estimated value should depend on the parent value. The smaller the parent value, the greater the shrinkage. This result is very interesting because it illustrates the effect of taking into account the parent-child dependency.

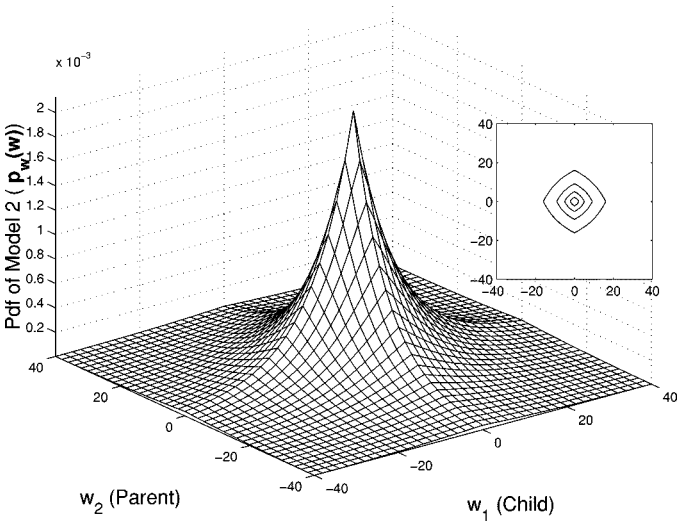

Fig. 9. New bivariate pdf (31) proposed for joint pdf of parent-child wavelet coefficient pairs (Model 2).

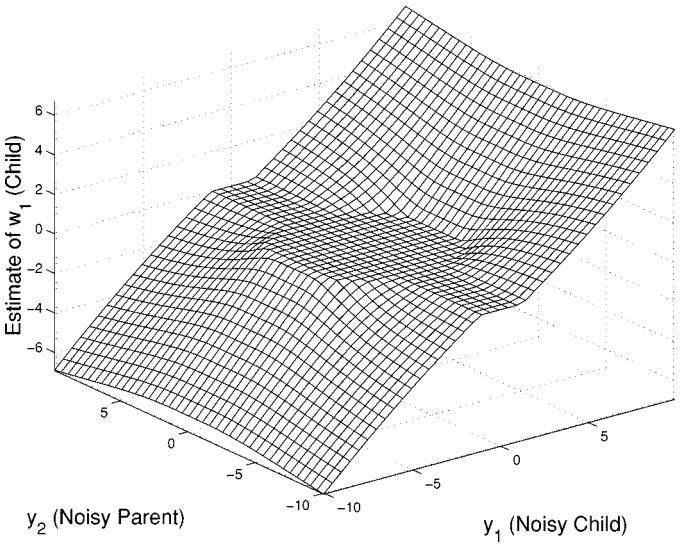

Fig. 10. New bivariate shrinkage function derived from the Model 2 proposed in (31) (Fig. 9).

Note that when the parent is zero, the MAP estimate of w_1 is obtained by the soft threshold function. If parent $y_2 = 0$, then $\hat{w}_1 = \text{soft}(y_1, \sqrt{3}\sigma_n^2/\sigma)$.

2) Model 2: Although Model 1 given in (21) is a better approximation to the empirical histogram than the independent Laplacian model, it is not possible to represent independent distributions with it. To characterize a larger group of probability distributions for the wavelet pairs, we need a more flexible model. For this purpose, we propose a pdf that can vary between Model 1 and the independent Laplacian model given in (22) with tunable parameters. The proposed joint pdf, called Model 2 in this paper, can be written as

$$p_{\mathbf{w}}(\mathbf{w}) = K \cdot \exp\left(-\left[a\sqrt{w_1^2 + w_2^2} + b(|w_1| + |w_2|)\right]\right) \tag{31}$$

where K is the normalization constant. A plot of this model is illustrated in Fig. 9. Let us develop the MAP estimator for this model. From (31)

$$f(\mathbf{w}) = -\left[a\sqrt{w_1^2 + w_2^2} + b(|w_1| + |w_2|)\right] \tag{32}$$

and this gives

$$f_1(\mathbf{w}) = -\left(\frac{a \cdot w_1}{\sqrt{w_1^2 + w_2^2}} + b \cdot \text{sign}(w_1)\right) \tag{33}$$

$$f_2(\mathbf{w}) = -\left(\frac{a \cdot w_2}{\sqrt{w_1^2 + w_2^2}} + b \cdot \text{sign}(w_2)\right). \tag{34}$$

Solving (25) and (26) using (33) and (34), the bivariate shrinkage function for this model can be obtained as

$$\hat{w}_1 = \frac{(R - \sigma_n^2 \cdot a)_+}{R} \cdot \text{soft}\left(y_1, \sigma_n^2 \cdot b\right) \tag{35}$$

where soft is defined in (14) and

$$R = \sqrt{\text{soft}(y_1, \sigma_n^2 \cdot b)^2 + \text{soft}(y_2, \sigma_n^2 \cdot b)^2}. \tag{36}$$

The derivation can be found in Appendix B. A plot of the bivariate shrinkage function using this model is illustrated in Fig. 10. As can be easily observed, there is a circular-like deadzone around the origin as in Model 1, but the deadzone also contains a stripe corresponding to large parent values, as in the independent Laplacian model. One can adjust the width of the stripe-shaped deadzone and the radius of the circular-like deadzone with the tunable parameters a and b. The bivariate shrinkage function derived from Model 2 specializes both to standard soft shrinkage function (Fig. 7) and to circular bivariate shrinkage function (Fig. 8).

Note that when $a = 0$, we recover the scalar soft thresholding rule; when $b = 0$, we recover the (30).

It should be noted that the selection of the model parameters a and b via maximum likelihood for Model 2 turns out to be more complicated than Model 1 because obtaining an expression for the parameters requires integrals that we cannot evaluate in closed form.

3) Model 3: In practice, the variance of the wavelet coefficients of natural images are quite different from scale to scale. We would like to generalize Model 1 since the marginal variances are the same. For this purpose, we propose Model 3, which has adjustable marginal variances, i.e.,

$$p_{\mathbf{w}}(\mathbf{w}) = \frac{3}{2\pi\sigma_1\sigma_2} \cdot \exp\left(-\sqrt{3} \cdot \sqrt{\left(\frac{w_1}{\sigma_1}\right)^2 + \left(\frac{w_2}{\sigma_2}\right)^2}\right). \tag{37}$$

Let us develop the MAP estimator for this model. From the pdf

$$f(\mathbf{w}) = -\sqrt{3} \cdot \sqrt{\left(\frac{w_1}{\sigma_1}\right)^2 + \left(\frac{w_2}{\sigma_2}\right)^2} \tag{38}$$

and this gives

$$f_1(\mathbf{w}) = -\frac{\sqrt{3}}{\sigma_1^2} \cdot \frac{w_1}{\sqrt{\left(\frac{w_1}{\sigma_1}\right)^2 + \left(\frac{w_2}{\sigma_2}\right)^2}} \tag{39}$$

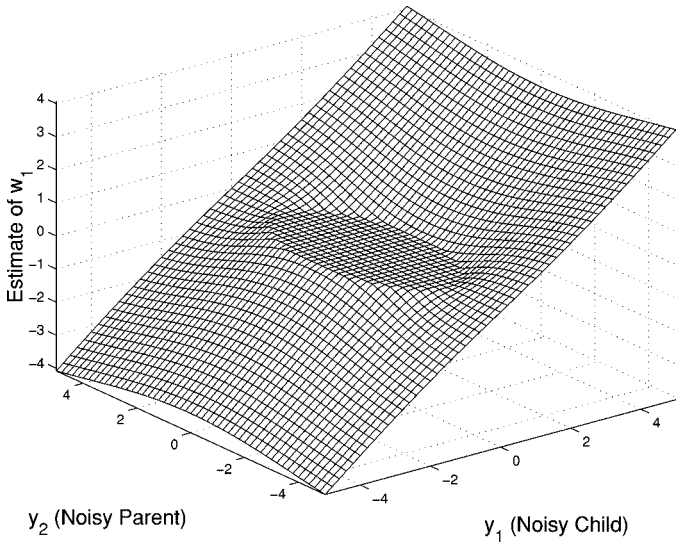

Fig. 11. Plot of the new bivariate shrinkage function derived from the Model 3 proposed in (37).

$$f_2(\mathbf{w}) = -\frac{\sqrt{3}}{\sigma_2^2} \cdot \frac{w_2}{\sqrt{\left(\frac{w_1}{\sigma_1}\right)^2 + \left(\frac{w_2}{\sigma_2}\right)^2}}. \tag{40}$$

Substituting (39) and (40) into the (25) and (26) gives

$$\hat{w}_1 \cdot \left(1 + \frac{\sqrt{3}\sigma_n^2}{\sigma_1^2 r}\right) = y_1 \tag{41}$$

$$\hat{w}_2 \cdot \left(1 + \frac{\sqrt{3}\sigma_n^2}{\sigma_2^2 r}\right) = y_2 \tag{42}$$

where

$$r = \sqrt{\left(\frac{\hat{w}_1}{\sigma_1}\right)^2 + \left(\frac{\hat{w}_2}{\sigma_2}\right)^2}. \tag{43}$$

These two equations do not have a simple closed-form solution like Model 1. This means that there is no clean expression for the bivariate shrinkage function. However, the solution can be found using iterative numerical methods. The solution using the successive substitution method and Newton–Raphson method will be described in Section II-B4 as a special case.

A plot of the bivariate shrinkage function using Model 3 is illustrated in Fig. 11. By means of the marginal variances σ_1 and σ_2 of Model 3, any ellipsoidal deadzone can be obtained. It does not restrict us to a circular deadzone in the shrinkage function, as in Model 1. Although it is not obvious, the deadzone can be written as

$$\text{deadzone} = \left\{(y_1, y_2): \sqrt{(y_1 \cdot \sigma_1)^2 + (y_2 \cdot \sigma_2)^2} \le \sqrt{3}\sigma_n^2\right\}.$$

Note that if $\sigma_1 = \sigma_2 = \sigma$, we get Model 1.

4) Model 4: Although Model 2 allows us to make a transition from the independent model and the newly proposed dependent model (Model 1), the marginal variances are the same for Model 2. Therefore, we would like to generalize Model 2. The proposed joint pdf is

$$p_{\mathbf{w}}(\mathbf{w}) = K \cdot \exp\left(-\left[\sqrt{c_1 \cdot w_1^2 + c_2 \cdot w_2^2} + c_3 \cdot |w_1| + c_4 \cdot |w_2|\right]\right) \tag{44}$$

where K is the normalization constant. Let us call it Model 4 and try to develop the MAP estimator for this model. From the pdf

$$f(\mathbf{w}) = -\left[\sqrt{c_1 \cdot w_1^2 + c_2 \cdot w_2^2} + c_3 \cdot |w_1| + c_4 \cdot |w_2|\right] \tag{45}$$

and it gives

$$f_1(\mathbf{w}) = -\left(\frac{c_1 \cdot w_1}{\sqrt{c_1 \cdot w_1^2 + c_2 \cdot w_2^2}} + c_3 \cdot \text{sign}(w_1)\right) \tag{46}$$

$$f_2(\mathbf{w}) = -\left(\frac{c_2 \cdot w_2}{\sqrt{c_1 \cdot w_1^2 + c_2 \cdot w_2^2}} + c_4 \cdot \text{sign}(w_2)\right). \tag{47}$$

Substituting (46) and (47) into the (25) and (26) gives

$$\hat{w}_1 \cdot \left(1 + \frac{c_1\sigma_n^2}{r}\right) = \text{soft}\left(y_1, c_3\sigma_n^2\right) \tag{48}$$

$$\hat{w}_2 \cdot \left(1 + \frac{c_2\sigma_n^2}{r}\right) = \text{soft}\left(y_2, c_4\sigma_n^2\right) \tag{49}$$

where

$$r = \sqrt{c_1 \cdot \hat{w}_1^2 + c_2 \cdot \hat{w}_2^2}. \tag{50}$$

These two equations also do not have a simple closed-form solution. This means there is no simple expression for the bivariate shrinkage function. Like Model 3, iterative numerical methods can be used to obtain the solution. The solution using the successive substitution method and Newton–Raphson method can be described as follows.

Algorithm Using the Successive Substitution Method: The algorithm can be given as follows.

1) Initialize $\hat{w}_1^{[0]}$ and $\hat{w}_2^{[0]}$, for example, $\hat{w}_1^{[0]} = y_1$ and $\hat{w}_2^{[0]} = y_2$, and $k = 0$.
2) Calculate r using $r = \sqrt{c_1 \cdot \left(\hat{w}_1^{[k]}\right)^2 + c_2 \cdot \left(\hat{w}_2^{[k]}\right)^2}$.
3) Find $\hat{w}_1^{[k+1]}$ and $\hat{w}_2^{[k+1]}$ using

$$\hat{w}_1^{[k+1]} = \frac{\text{soft}\left(y_1, c_3\sigma_n^2\right)}{\left(1 + \frac{c_1\sigma_n^2}{r}\right)} \tag{51}$$

$$\hat{w}_2^{[k+1]} = \frac{\text{soft}\left(y_2, c_4\sigma_n^2\right)}{\left(1 + \frac{c_2\sigma_n^2}{r}\right)}. \tag{52}$$

4) Find the differences, i.e., $\epsilon_1 = \hat{w}_1^{[k+1]} - \hat{w}_1^{[k]}$ and $\epsilon_2 = \hat{w}_2^{[k+1]} - \hat{w}_2^{[k]}$.

5) If both ϵ_1 and ϵ_2 are small, then terminate the iteration. Otherwise, set $k = k+1$, go to step 2.

Algorithm Using Newton Raphson Method: The convergence is only linear if the simple successive iteration method is used. To improve the rate of convergence, the Newton-Raphson method, which has quadratic convergence, can be used. The general solution of this method for (48) and (49) requires the solution over two variables, namely, w_1 and w_2. However, to reduce the computational complexity, we describe a modification that transforms the problem into one having one variable. In Appendix C, it is proven that it is guaranteed to converge. Using (48) and (49) with (50), one gets

$$g(r) := c_1 \cdot \frac{\text{soft}\left(y_1, c_3\sigma_n^2\right)^2}{\left(r + c_1\sigma_n^2\right)^2} + c_2 \cdot \frac{\text{soft}\left(y_2, c_4\sigma_n^2\right)^2}{\left(r + c_2\sigma_n^2\right)^2} - 1 = 0. \quad (53)$$

Therefore, our problem reduces to finding a solution for $g(r) = 0$ using the Newton-Raphson method. After r is obtained, the solution can be found using (48) and (49).

Newton's iteration can be stated as

$$r^{[k+1]} = r^{[k]} - \frac{g\left(r^{[k]}\right)}{g'\left(r^{[k]}\right)}. \quad (54)$$

From (53)

$$g'(r) = -2\left(c_1 \cdot \frac{\text{soft}\left(y_1, c_3\sigma_n^2\right)^2}{\left(r + c_1\sigma_n^2\right)^3} + c_2 \cdot \frac{\text{soft}(y_2, c_4\sigma_n^2)^2}{(r + c_2\sigma_n^2)^3}\right). \quad (55)$$

The Newton–Raphson algorithm can be given step-by-step as follows.

1) Initiliaze $r^{[0]}$, for example, $r^{[0]} = c_1 y_1^2 + c_2 y_2^2$ and $k = 0$.
2) Calculate $g\left(r^{[k]}\right)$ using (53).
3) Calculate $g'\left(r^{[k]}\right)$ using (55).
4) Calculate $r^{[k+1]}$ using (54).
5) Find the difference between $r^{[k+1]}$ and $r^{[k]}$, $\epsilon = r^{[k+1]} - r^{[k]}$
6) If ϵ is small, go step 8. Otherwise, $k = k+1$, and o to step 2.
7) Using the $r^{[k+1]}$ calculated after this iterartion, w_1 can be obtained as

$$\hat{w}_1 = \frac{\text{soft}\left(y_1, c_3\sigma_n^2\right)}{1 + \frac{c_1\sigma_n^2}{r^{[k+1]}}}. \quad (56)$$

A plot of a bivariate shrinkage function using Model 4 is illustrated in Fig. 12. By means of the tunable parameters, c_1, c_2, c_3, and c_4, the Model 4 does not restrict us to a circular deadzone in the shrinkage function as in the Models 1 and 2. Using the parameters, any kind of ellipsoidal deadzone can be obtained, and the width of the stripe deadzone can be adjustable. Note that if $c_1 = c_2 = a^2$ and $c_3 = c_4 = b$, we get Model 2.

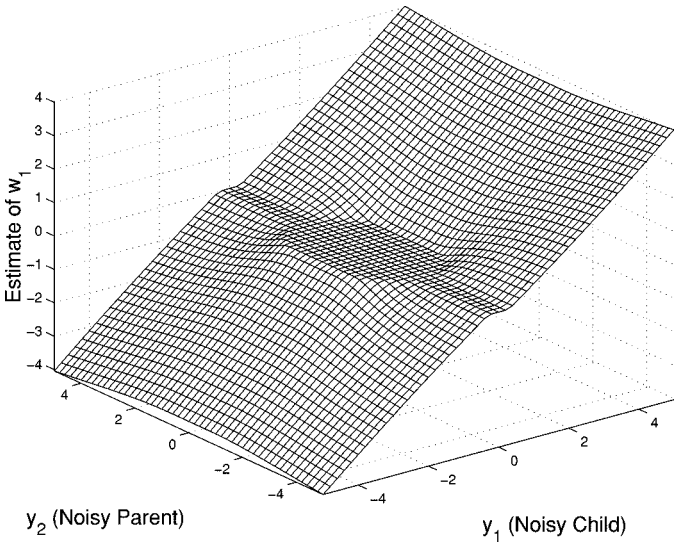

Fig. 12. Plot of the new bivariate shrinkage function derived from the Model 4 proposed in (44).

III. Application to Image Denoising

In Section II, we have proposed new joint statistical models for wavelet coefficients and obtained MAP estimators for each model (in two cases, in closed form). This section presents three image denoising examples to show the efficiency of these new models and compares them with other methods in the literature. In the first example, Model 1 will be used and will be compared with classical soft thresholding. In the second example, we will develop a subband dependent data-driven image denoising algorithm similar to BayesShrink, which uses Model 3, and the results will be compared with VisuShrink, SureShrink, BayesShrink, and the hidden Markov tree model, which are also data-driven systems. In our numerical experiments, it was found that Model 2 and 4 give negligible improvement in image denoising. (In those experiments, the optimal parameter values were found using a search.) Therefore, the selection of the tunable parameters in Model 2 and 4 was not developed. In addition, Model 3 gave marginal improvement over Model 1, but for Model 3, the parameters σ_1 and σ_2 can be found in the way described in Section III-A. In the third example, the performance of a subband-dependent system using Model 1 will be demonstrated on the dual-tree complex wavelet transform.

A. Example 1

In this experiment, a critically sampled orthogonal discrete wavelet transform, with Daubechies length-8 filter, is used. We have compared our bivariate shrinkage function (30) with the classical soft thresholding estimator given in (12) for image denoising. The 512×512 Lena image is used for this purpose. Zero mean white Gaussian noise is added to the original image ($\sigma_n = 25.5$). The PSNR value of the noisy image is 20.02 dB. Part of the original and the noisy images are illustrated in Fig. 13(a) and (b).

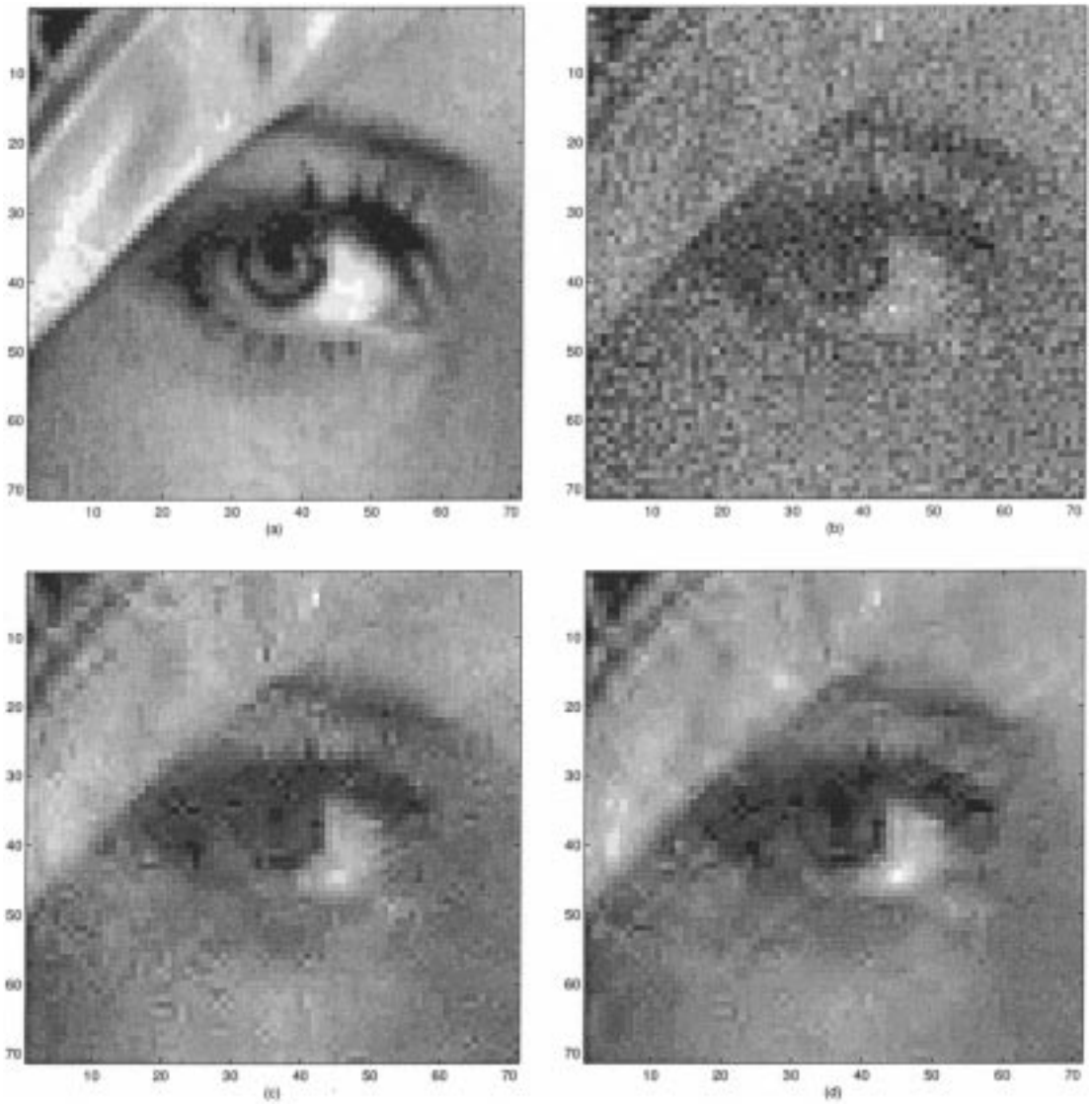

Fig. 13. (a) Original image. (b) Noisy image with PSNR $= 20.02$ dB. (c) Denoised image using soft thresholding; PSNR $= 27.73$ dB. (d) Denoised image using new bivariate shrinkage function given in (30); PSNR $= 28.83$ dB.

The denoised image obtained using the soft threshold has a PSNR of 27.73 dB [Fig. 13(c)]. The denoised image obtained using the new bivariate shrinkage function has a PSNR of 28.83 dB (Fig. 13(d)). In addition, scalar hard thresholding results in a PSNR of 27.54 dB (not shown). The threshold value in each case was chosen to maximize the PSNR.

B. Example 2

In this example, a subband adaptive data-driven image denoising system that uses Model 3 given will be described, and some comparison to VisuShrink [14], SureShrink [15], BayesShrink [4], and the hidden Markov tree model [11] will be given. Our system exploits only the parent-child dependencies as opposed to the ones exploiting intra-scale dependencies [29], [30], [36], [40].

In this system, the wavelet pairs in each subband are assumed to be realizations from Model 3 to create a system that is adaptive to different subband characteristics. As described in Section II-B3, the solution for the MAP estimator needs iterative methods. In addition, the solution requires the *a priori* knowledge of the noise variance σ_n^2 and the marginal variances σ_1^2 and σ_2^2 for a data-driven system. In our system, the marginal variances are calculated separately for each subband in order to have subband adaptivity.

Fig. 14 illustrates the subband regions of the two-dimensional (2-D) critically sampled wavelet transform. For convenience, let us label the subbands HH_k, HL_k, and LH_k, where k is the scale, and J is the coarsest scale. The smaller k is, the finer the scale is. Let us also define subband $P(S)$. $P(S)$ is the subband of the parents of the coefficients of the subband S. For example, if S is HH_1, then $P(S)$ is HH_2, or if S is HL_2, then $P(S)$ is HL_3.

To estimate the noise variance σ_n^2 from the noisy wavelet coefficients, a robust median estimator is used from the finest scale wavelet coefficients (HH_1 subband) [14].

$$\hat{\sigma}_n^2 = \frac{\text{median}(|y_i|)}{0.6745}, \quad y_i \in \text{ subband } HH_1. \tag{57}$$

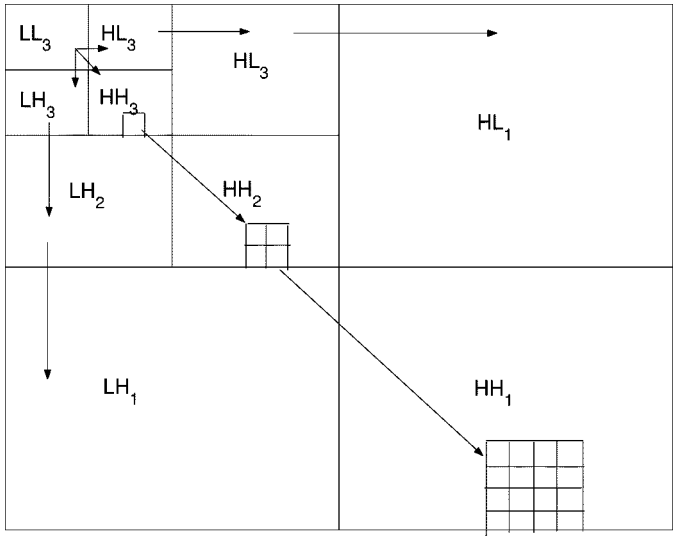

Fig. 14. Subband regions of critically sampled wavelet transform.

Let us assume that we are trying to estimate the marginal variances σ_1 and σ_2 for the subbands S and $P(S)$. Recall our observation model

$$y_1 = w_1 + n_1$$
$$y_2 = w_2 + n_2$$

where y_1, w_1, $n_1 \in S$, and y_2, w_2, $n_2 \in P(S)$. Since w_1 and n_1 and w_2 and n_2 are independent of each other, one gets

$$\sigma_{y_1}^2 = \sigma_1^2 + \sigma_n^2 \tag{58}$$
$$\sigma_{y_2}^2 = \sigma_2^2 + \sigma_n^2 \tag{59}$$

where σ_{y_1} and σ_{y_2} are the variances of y_1 and y_2. Since y_1 and y_2 are modeled as zero mean, σ_{y_1} and σ_{y_2} can be found empirically by

$$\hat{\sigma}_{y_1}^2 = \frac{1}{N_1^2} \sum_{y_{1i} \in S} y_{1i}^2 \tag{60}$$
$$\hat{\sigma}_{y_2}^2 = \frac{1}{N_2^2} \sum_{y_{2i} \in P(S)} y_{2i}^2 \tag{61}$$

where N_1^2 and N_2^2 are the sizes of the subbands S and $P(S)$, respectively. Although these are empirical, the same results can be obtained with maximum likelihood (ML) estimator of σ_{y_i} if one assumes y_1 and y_2 are Gaussian and uses N_1^2 observations for y_1 and N_2^2 observations for y_2 in order to estimate σ_{y_1} and σ_{y_2}. However, if y_1 and y_2 are assumed to be Laplacian, the ML estimator of σ_{y_i} is given by

$$\hat{\sigma}_{y_1} = \frac{\sqrt{2}}{N_1^2} \sum_{y_{1i} \in S} |y_{1i}| \tag{62}$$
$$\hat{\sigma}_{y_2} = \frac{\sqrt{2}}{N_2^2} \sum_{y_{2i} \in P(S)} |y_{2i}|. \tag{63}$$

Either of the equation pairs (60) and (61) or (62) and (63) can be used as the estimates of σ_{y_1} and σ_{y_2}. In our experiments, we obtained better PSNR values with our model if we use the Laplacian assumption. Therefore, in our system, we use (62) and (63).

Using (58) and (59), σ_1 and σ_2 can be estimated as

$$\hat{\sigma}_1 = \sqrt{(\hat{\sigma}_{y_1}^2 - \hat{\sigma}_n^2)_+} \tag{64}$$
$$\hat{\sigma}_2 = \sqrt{(\hat{\sigma}_{y_2}^2 - \hat{\sigma}_n^2)_+}. \tag{65}$$

Now, everything that is necessary in order to apply a MAP estimator corresponding to Model 3 $(\sigma_1, \sigma_2, \sigma_n)$ is estimated. Either the successive substitution method or the Newton–Raphson method described in Section II-B3 can be used to estimate wavelet coefficients. This algorithm results in coefficient and parent estimates. We only use coefficient estimates. For simplicity, we did not exploit the double estimation of coefficients above the finest scale.

Let us summarize the algorithm.

1) Calculate the noise variance $\hat{\sigma}_n^2$ using (57).
2) For each subband, $(S = HH_i, HL_i, LH_i, i = 1, \ldots, J)$.
 a) Calculate $\hat{\sigma}_{y_1}$ and $\hat{\sigma}_{y_2}$ using (62) and (63);
 b) Calculate $\hat{\sigma}_1$ and $\hat{\sigma}_2$ using(64) and (65);
 c) Estimate each coefficient using either the successive substitution method or the Newton–Raphson method described in Section II-B3.

In this experiment we used three 512×512 grayscale images, namely, Lena, Boat, and Barbara. This algorithm was tested using different noise levels σ_n =10, 20, and 30 and compared with VisuShrink, SureShrink, BayesShrink, and HMT. Performance analysis is done using the PSNR measure. Let s, d denote the original and the denoisied image. The rms error is given by

$$\epsilon = \sqrt{\frac{1}{N^2} \sum_k (s_k - d_k)^2} \tag{66}$$

where N^2 is the number of pixels. The PSNR in decibels is given by

$$\text{PSNR} = 20 \log_{10} \left(\frac{256}{\epsilon} \right). \tag{67}$$

Each PSNR value in the table is averaged over five runs. The results can be seen in Table I. In this table, the highest PSNR value among three algorithms is emphasized with a star $(\star)$. As seen from the results, our algorithm mostly outperfoms the others.

Other image denoising techniques that exploit *intrascale* dependencies [26], [29], [30], [41] yield better performance than the proposed algorithm does. We are currently investigating extensions of the proposed algorithm in order to exploit intrascale dependencies.

TABLE I
AVERAGE PSNR VALUES OF DENOISED IMAGES OVER FIVE RUNS FOR DIFFERENT TEST IMAGES AND NOISE LEVELS (σ_n) OF NOISY, VISUSHRINK, SURESHRINK, BAYESSHRINK, HMT SYSTEM, AND OUR SYSTEM DESCRIBED IN SECTION III-B

	Noisy	VisuShrink	SureShrink	BayesShrink	HMT	Our Model 3	Dual Tree
Lena							
$\sigma = 10$	28.18	28.76	33.28	33.32	33.84	33.94 $\star$	34.77
$\sigma = 20$	22.14	26.46	30.22	30.17	30.39	30.73 $\star$	31.71
$\sigma = 30$	18.62	25.14	28.38	28.48	28.35	28.94 $\star$	29.85
Boat							
$\sigma = 10$	28.16	26.49	31.19	31.80	32.28 $\star$	32.25	32.85
$\sigma = 20$	22.15	24.43	28.14	28.48	28.84	28.93 $\star$	29.81
$\sigma = 30$	18.62	23.33	26.52	26.60	26.83	27.11 $\star$	27.99
Barbara							
$\sigma = 10$	28.16	24.81	30.21	30.86	31.36 $\star$	31.13	31.30
$\sigma = 20$	22.14	22.81	25.91	27.13	27.80 $\star$	27.25	27.78
$\sigma = 30$	18.62	22.00	24.33	25.16	25.11	25.21 $\star$	25.62

C. Example 3

In this example, we will demonstrate the performance of our bivariate shrinkage function derived from Model 1 on the dual-tree complex wavelet transform [23], [24], and the performance will be tested with a subband adaptive denoising system like the one described in Example 2.

The dual-tree DWT is an overcomplete wavelet transform, which can be implemented by two wavelet filterbanks operating in parallel. The performance gains provided by the dual-tree DWT come from designing the filters in the two filter banks appropriately. The coefficients produced by these filterbanks are the real and imaginary parts of a complex coefficient. Assume the sets of coefficients u_i and v_i are produced by these filterbanks separately, and the complex coefficients can represented by $c_i = u_i + v_i$.

The properties of dual-tree DWT include the following.

- It is nearly shift invariant, i.e., small signal shifts do not affect the magnitudes of the complex coefficients $\left(|c_i| = \sqrt{u_i^2 + v_i^2}\right)$, although they do affect the real and imaginary parts. Therefore, the magnitude information is a more reliable measure than either the real u_i or the imaginary v_i parts.

1) The basis functions have directional selectivity property at ± 15, ± 45, and $\pm 75^\circ$, which the regular critically sampled transform does not have.
2) For m-dimensional signals, it has 2^m times redundancy, for example, four times redundant for images.

The new bivariate shrinkage function will be applied to the magnitude of the dual-tree DWT coefficients since it is more shift invariant than the real or imaginary parts. We assume the that magnitudes of the coefficients are corrupted by additive Gaussian noise, even though they are not.

The performance of this system is tested with the same experiment in Example 2. The PSNR values are illustrated in the last column of Table I. From this table, it is evident that using our bivariate shrinkage function with the dual-tree DWT provides better performance than using it with the critically sampled DWT. In [8], the HMT modeling is also extended to the dual-tree DWT (CHMT). Our experiments suggest that for high noise levels, the bivariate shrinkage procedure described here can be competitive with the CHMT.

IV. CONCLUSION AND FUTURE WORK

In this paper, first four new bivariate distributions are proposed for wavelet coefficients of natural images in order to characterize the dependencies between a coefficient and its parent, and second, the corresponding bivariate shrinkage functions are derived from them using Bayesian estimation, in particular, the MAP estimator. Two of these new bivariate shrinkage functions (Model 1 and 2) are given by simple formulas. Therefore, they maintain the simplicity, efficiency, and intuition of the classical soft thresholding approach. In order to characterize larger group of distributions, Models 3 and 4 are proposed, and numerical solutions for the MAP estimators are given and are proven to converge.

In order to show the effectiveness of these new estimators, three examples are presented and compared with effective techniques in the literature. In the second example, a subband-adaptive data-driven system is developed and compared with the HMT model [11], which exploits the interscale dependencies of coefficients and BayesShrink [4], which is also a subband-adaptive data-driven system, which outperforms VisuShrink and SureShrink. In our experiments, our system mostly outperforms the others. The performance of a subband-adaptive data-driven system is also demonstrated on the dual-tree complex wavelet transform as another example.

It should be emphasized that in this paper, we investigate only how the classical soft thresholding approach of Donoho and Johnstone [14] should be modified to take into account parent-child statistics. State-of-the-art denoising algorithms [3], [8], [29], [30] generally use local adaptive methods or in other ways exploit dependencies between larger numbers of coefficients. Using local adaptive methods in combination with bivariate shrinkage may further improve the denoising results reported in Section III. Our experiments showed that the use of our models 2, 3, and 4 resulted in negligible improvement on image denoising performance over our Model 1. Therefore, in practice we suggest Model 1 due to its simplicity and efficiency. Other

simple bivariate shrinkage functions can also be developed, for example, a bivariate hard threshold with a circular or ellipsoidal deadzone, or a bivariate generalization of the semi-soft rule of [17] and [18].

We obtained these results by observing the dependencies between only coefficients and their parents. It is expected that the results can be further improved if the other dependencies between a coefficient and its other neighbors are exploited. Hence, we are currently investigating multivariate extensions of this new bivariate shrinkage rule.

Appendix A
Derivation of the Shrinkage Function for Model 1

Substituting (28) and (29) into the (25) and (26) gives

$$\hat{w}_1 \cdot \left(1 + \frac{\sqrt{3}\sigma_n^2}{\sigma r}\right) = y_1, \quad \hat{w}_2 \cdot \left(1 + \frac{\sqrt{3}\sigma_n^2}{\sigma r}\right) = y_2 \tag{68}$$

where $r = \sqrt{\hat{w}_1^2 + \hat{w}_2^2}$. Using (68)

$$\begin{aligned} r^2 =& \frac{y_1^2}{\left(1 + \frac{\sqrt{3}\sigma_n^2}{\sigma r}\right)^2} + \frac{y_2^2}{\left(1 + \frac{\sqrt{3}\sigma_n^2}{\sigma r}\right)^2} \\ \left(r + \frac{\sqrt{3}\sigma_n^2}{\sigma}\right)^2 =& y_1^2 + y_2^2 \\ r =& \left(\sqrt{y_1^2 + y_2^2} - \frac{\sqrt{3}\sigma_n^2}{\sigma}\right)_+ . \end{aligned} \tag{69}$$

Substituting r in (68) gives

$$\hat{w}_1 = \frac{\left(\sqrt{y_1^2 + y_2^2} - \frac{\sqrt{3}\sigma_n^2}{\sigma}\right)_+}{\sqrt{y_1^2 + y_2^2}} \cdot y_1. \tag{70}$$

Appendix B
Derivation of the Shrinkage Function for Model 2

Substituting (33) and (34) into (25) and (26) gives

$$\begin{aligned} \hat{w}_1 \cdot \left(1 + \frac{a\sigma_n^2}{r}\right) &= \left(|y_1| - b\sigma_n^2\right)_+ \cdot \operatorname{sign}(y_1) \\ &= \operatorname{soft}\left(y_1, b\sigma_n^2\right) \end{aligned} \tag{71}$$

$$\begin{aligned} \hat{w}_2 \cdot \left(1 + \frac{a\sigma_n^2}{r}\right) &= \left(|y_2| - b\sigma_n^2\right)_+ \cdot \operatorname{sign}(y_2) \\ &= \operatorname{soft}\left(y_2, b\sigma_n^2\right) \end{aligned} \tag{72}$$

where $r = \sqrt{\hat{w}_1^2 + \hat{w}_2^2}$. Using (71) and (72)

$$\begin{aligned} r^2 =& \frac{\operatorname{soft}\left(y_1, b\sigma_n^2\right)^2}{\left(1 + \frac{a\sigma_n^2}{r}\right)^2} + \frac{\operatorname{soft}\left(y_2, b\sigma_n^2\right)^2}{\left(1 + \frac{a\sigma_n^2}{r}\right)^2} \\ \left(r + a\sigma_n^2\right)^2 =& \operatorname{soft}\left(y_1, b\sigma_n^2\right)^2 + \operatorname{soft}\left(y_2, b\sigma_n^2\right)^2 \\ r =& \left(\sqrt{\operatorname{soft}(y_1, b\sigma_n^2)^2 + \operatorname{soft}(y_2, b\sigma_n^2)^2} - a\sigma_n^2\right)_+ \\ r =& \left(R - a\sigma_n^2\right)_+ \end{aligned} \tag{73}$$

where $R = \sqrt{\operatorname{soft}(y_1, b\sigma_n^2)^2 + \operatorname{soft}(y_2, b\sigma_n^2)^2}$. Substituting r in (71) gives

$$\hat{w}_1 = \frac{\left(R - \sigma_n^2 \cdot a\right)_+}{R} \cdot \operatorname{soft}\left(y_1, \sigma_n^2 \cdot b\right). \tag{74}$$

Appendix C
Proof of Convergence

In this section, we will prove that Newton's methods described in Sections II-B3 and 4 for Model 3 and Model 4 are convergent for all initial conditions. Since Model 3 is special case of Model 4, we will examine Model 4. Our problem is to find the r value where

$$\begin{aligned} g(r) = c_1 \cdot \frac{\operatorname{soft}\left(y_1, c_3\sigma_n^2\right)^2}{\left(r + c_1\sigma_n^2\right)^2} \\ + c_2 \cdot \frac{\operatorname{soft}\left(y_2, c_4\sigma_n^2\right)^2}{\left(r + c_2\sigma_n^2\right)^2} - 1 = 0. \end{aligned} \tag{75}$$

Since r is defined as

$$r = \sqrt{c_1 \cdot w_1^2 + c_2 \cdot w_2^2} \tag{76}$$

r is always greater than zero, and $r \geq 0$, i.e., $g(r)$ is defined only on $r \geq 0$. If we take the derivative of $g(r)$, we get

$$g'(r) = -2\left(c_1 \cdot \frac{\operatorname{soft}(y_1, c_3\sigma_n^2)^2}{(r + c_1\sigma_n^2)^3} + c_2 \cdot \frac{\operatorname{soft}(y_2, c_4\sigma_n^2)^2}{(r + c_2\sigma_n^2)^3}\right). \tag{77}$$

Note that $g'(r) < 0$ for all r values since $c_1 > 0$, $c_2 > 0$, and $r > 0$. Therefore, $g(r)$ is a decreasing function, which means that $g(r)$ has a maximum at $r = 0$. Then, if $g(r)$ is positive, it has a zero, and the zero is unique. Besides, if $g(r)$ is negative, $g(r)$ does not have a zero, which means Newton's iteration does not have a solution, but $r = 0$ maximizes the MAP estimator. Therefore, one can assume that $r = 0$ is the solution for the Newton's iteration, i.e., set $r = 0$ as a zero for $g(r) = 0$ if $g(0) < 0$.

Let us consider the case $g(r) > 0$. The second derivative of $g(r)$ can be written as

$$g''(r) = 6\left(c_1 \cdot \frac{\operatorname{soft}(y_1, c_3\sigma_n^2)^2}{(r + c_1\sigma_n^2)^4} + c_2 \cdot \frac{\operatorname{soft}(y_2, c_4\sigma_n^2)^2}{(r + c_2\sigma_n^2)^4}\right). \tag{78}$$

From this, it can be concluded that $g''(r) > 0$ for all r values, which means that $g(r)$ is a convex function. Therefore, if a function is convex and has a unique zero, the Newton iteration will converge to it from any starting point. In our case, we need a small modification since the function is defined only for $r \geq 0$ values. If an iteration gives $r < 0$ values, then set $r = 0$.

References

[1] F. Abramovich and Y. Benjamini, "Adaptive thresholding of wavelet coefficients," *Comput. Statist. Data Anal.*, vol. 22, pp. 351–361, 1996.

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

[3] Z. Cai, T. H. Cheng, C. Lu, and K. R. Subramanian, "Efficient wavelet-based image denoising algorithm," *Electron. Lett.*, vol. 37, no. 11, pp. 683–685, May 2001.

[4] S. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Trans. Image Processing*, vol. 9, pp. 1532–1546, Sept. 2000.
[5] S. G. Chang, B. Yu, and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," *IEEE Trans. Image Processing*, vol. 9, pp. 1522–1531, Sept. 2000.
[6] H. Choi and R. Baraniuk, "Multiscale texture segmentation using wavelet-domain hidden Markov models," in *Proc. Int. Conf. Signals, Syst., Comput.*, vol. 2, 1998, pp. 1692–1697.
[7] H. Choi and R. G. Baraniuk, "Wavelet statistical models and Besov spaces," in *Proc. SPIE Tech. Conf. Wavelet Applicat. Signal Process.*, July 1999.
[8] H. Choi, J. K. Romberg, R. G. Baraniuk, and N. G. Kingsbury, "Hidden Markov tree modeling of complex wavelet transforms," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, vol. 1, Istanbul, Turkey, June 2000, pp. 133–136.
[9] C. Christopoulos, A. Skodras, and T. Ebrahimi, "The JPEG2000 still image coding system: An overview," *IEEE Trans. Consum. Electron.*, vol. 46, pp. 1103–1127, July 1993.
[10] R. Coifman and D. Donoho, "Time-invariant wavelet denoising," in *Wavelet and Statistics*, A. Antoniadis and G. Oppenheim, Eds. New York: Springer-Verlag, 1995, vol. 103, Lecture Notes in Statistics, pp. 125–150.
[11] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based signal processing using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 46, pp. 886–902, Apr. 1998.
[12] J.-C. Pesquet and D. Leporini, "Bayesian wavelet denoising: Besov priors and non-Gaussian noises," *Signal Process.*, vol. 81, pp. 55–66, 2001.
[13] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inform. Theory*, vol. 41, pp. 613–627, May 1995.
[14] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.
[15] ——, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, no. 432, pp. 1200–1224, 1995.
[16] D. 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.
[17] M. A. T. Figueiredo and R. D. 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.
[18] H. Gao, "Wavelet shrinkage denoising using the nonnegative garrote," *J. Comput. Graph. Stat.*, vol. 7, pp. 469–488, 1998.
[19] J. Huang, "Statistics of Natural Images and Models," Ph.D. dissertation, Brown Univ., Providence, RI, 2000.
[20] J. Huang, A. Lee, and D. Mumford, "Statistics of range images," in *Proc. Conf. Comput. Vision Pattern Recogn.*, Hilton Head, SC, 2000.
[21] A. Hyvarinen, "Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation," *Neural Comput.*, vol. 11, pp. 1739–1768, 1999.
[22] A. Hyvarinen, E. Oja, and P. Hoyer, "Image denoising by sparse code shrinkage," in *Intelligent Signal Processing*, S. Haykin and B. Kosko, Eds. Piscataway, NJ: IEEE, 2001.
[23] N. G. Kingsbury, "Image processing with complex wavelets," *Phil. Trans. R. Soc. London A*, Sept. 1999.
[24] ——, "Complex wavelets for shift invariant analysis and filtering of signals," *Applied Comptat. Harmon. Anal.*, pp. 234–253, May 2001.
[25] J. Li and R. Gray, "Text and picture segmentation by the distribution analysis of wavelet coefficients," in *Proc. Int. Conf. Image Processing*, Chicago, Oct. 1998.
[26] X. Li and M. T. Orchard, "Spatially adaptive image denosing under overcomplete expansion," in *Proc. IEEE Int. Conf. Image Process.*, Sept. 2000.
[27] J. Liu and P. Moulin, "Image denoising based on scale-space mixture modeling of wavelet coefficients," in *Proc. IEEE Int. Conf. Image Process.*, Kobe, Japan, Oct. 1999.
[28] ——, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Trans. Image Processing*, vol. 10, pp. 1647–1658, Nov. 2001.
[29] M. K. Mihcak, I. Kozintsev, K. Ramchandran, and P. Moulin, "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Signal Processing Lett.*, vol. 6, pp. 300–303, Dec. 1999.
[30] J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, "Adaptive wiener denoising using a Gaussian scale mixture model," in *Proc. Int. Conf. Image Process.*, 2001.
[31] M. Rangaswamy, D. Weiner, and A. Ozturk, "Non-Gaussian random vector identification using spherically invariant random processes," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 29, pp. 111–123, Jan. 1993.
[32] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.
[33] E. Simoncelli, "Statistical models for images: Compression, restoration and synthesis," in *Proc. 31st Asilomar Conf. Signals, Syst., Comput.*, Nov. 1997, pp. 673–678.
[34] E. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," *Annu. Rev. Neurosci.*, vol. 24, pp. 1193–1216, May 2001.
[35] 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.
[36] ——, "Modeling the joint statistics of images in the wavelet domain," *Proc. SPIE*, vol. 313, no. 1, pp. 188–195, 1999.
[37] E. P. Simoncelli and E. H. Adelson, "Noise removal via bayesian wavelet coring," in *Proc. IEEE Int. Conf. Image Process.*, vol. I, Jan. 1996, pp. 379–382.
[38] V. Strela, J. Portilla, and E. Simoncelli, "Image denoising using a local Gaussian scale mixture model in the wavelet domain," in Proc. SPIE 45th Annu. Meet., 2000.
[39] B. Vidakovic, *Statistical Modeling by Wavelets*. New York: Wiley, 1999.
[40] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," *Adv. Neural Inform. Process. Syst.*, vol. 12, May 2000.
[41] G. Fan and X. G. Xia, "Image denoising using a local contextual hidden Markov model in the wavelet domain," *IEEE Signal Processing Lett.*, vol. 8, pp. 125–128, May 2001.
[42] ——, "Improved hidden Markov models in the wavelet-domain," *IEEE Trans. Signal Processing*, vol. 49, pp. 115–120, Jan. 2001.
[43] K. Yao, "A representation theorem and its applications to spherically-invariant random processes," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 600–608, Sept. 1973.
[44] T. Cai and B. W. Silverman, "Incorporating information on neighboring coefficients into wavelet esstimation," *Sankhya*, vol. 63, pp. 127–148, 2001.

Levent Şendur (S'00) received the B.S. and M.S. degrees in electrical engineering in 1996 and 1999, respectively, from Middle East Technical University, Ankara, Turkey. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Polytechnic University, Brooklyn, NY.

Ivan W. Selesnick (M'95) received the B.S., M.E.E., and Ph.D. degrees in electrical engineering from Rice University, Houston, TX, in 1990, 1991, and 1996, respectively, .

As a Ph.D. student, he received a DARPA-NDSEG fellowship in 1991. In 1997, he was a Visiting Professor at the University of Erlangen-Nurnberg, Germany. Since 1997, he has been an Assistant Professor with the Department of Electrical and Computer Engineering, Polytechnic University, Brooklyn, NY. His current research interests are in the areas of digital signal processing and wavelet-based signal processing.

Dr. Selesnick's Ph.D. dissertation received the Budd Award for Best Engineering Thesis at Rice University in 1996 and an award from the Rice-TMC chapter of Sigma Xi. In 1997, he received an Alexander von Humboldt Award. He received a National Science Foundation Career award in 1999. He is currently a member of the IEEE Signal Processing Theory and Methods Technical Committee and an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING.