Multiscale LMMSE-Based Image Denoising With Optimal Wavelet Selection

Lei Zhang, Paul Bao, *Senior Member, IEEE*, and Xiaolin Wu, *Senior Member, IEEE*

***Abstract*—In this paper, a wavelet-based multiscale linear minimum mean square-error estimation (LMMSE) scheme for image denoising is proposed, and the determination of the optimal wavelet basis with respect to the proposed scheme is also discussed. The overcomplete wavelet expansion (OWE), which is more effective than the orthogonal wavelet transform (OWT) in noise reduction, is used. To explore the strong interscale dependencies of OWE, we combine the pixels at the same spatial location across scales as a vector and apply LMMSE to the vector. Compared with the LMMSE within each scale, the interscale model exploits the dependency information distributed at adjacent scales. The performance of the proposed scheme is dependent on the selection of the wavelet bases. Two criteria, the signal information extraction criterion and the distribution error criterion, are proposed to measure the denoising performance. The optimal wavelet that achieves the best tradeoff between the two criteria can be determined from a library of wavelet bases. To estimate the wavelet coefficient statistics precisely and adaptively, we classify the wavelet coefficients into different clusters by context modeling, which exploits the wavelet intrascale dependency and yields a local discrimination of images. Experiments show that the proposed scheme outperforms some existing denoising methods.**

***Index Terms*—Context modeling, image denoising, multiresolution analysis, mutual information, optimal basis, wavelets.**

I. Introduction

STATISTICAL modeling is of essence for the effectiveness of signal processing. As a Karhunen–Loève like expansion, wavelet transform (WT) [1]–[5] can decorrelate random processes into nearly independent coefficients [6], which can then be more effectively modeled statistically. WT has been successfully applied to coding and denoising. Since the first wavelet soft thresholding approach of Donoho [9], many wavelet-based denoising schemes were reported [7]–[14], [16]–[18], [26], [28], [29].

WT packs most of the signal energy into a few significant coefficients and relates the insignificant coefficients to the signal-independent additive noise. In threshold-based denoising schemes, a threshold is set to distinguish noise from the structural information. Thresholding can be classified into soft and hard ones, in which coefficients less than the threshold will be set to 0 but those above the threshold will be preserved (hard thresholding) or shrunk (soft thresholding). Donoho [9] first presented the *WaveletShrinkage* scheme $\eta_t(w) = \mathrm{sgn}(w) \cdot \max(|w| - t, 0)$ with a universal threshold $t = \sigma\sqrt{2\log N}$ based on orthonormal wavelet bases, where w is the wavelet coefficient, σ is the noise standard deviation, and N is the sample length of signal. The threshold is claimed asymptotically optimal in minimax sense but it would over-smooth signals in practice. Since Donoho's pioneer work, a numerous threshold-based denoising schemes have been proposed [7], [8], [10], [11], [13], [14], [16]. It is generally accepted that in each subband the image wavelet coefficients can be modeled as independent identically distributed (i.i.d.) random variables with generalized Gaussian distribution (GGD) [3], [7], [8], with which Chang [7] presented a near optimal soft threshold $t = \sigma^2/\sigma_{W_j}$ (the wavelet base is assumed orthonormal), where σ_{W_j} is the standard deviation of wavelet coefficients at scale j. It reportedly outperformed that of the classical nonlinear *WaveletShrinkage* [9] and the improved *SureShrink* [10] of Donoho. The aforementioned three thresholds are soft, meaning that the input w would be shrunk to zero by an amount of threshold t, and derived with orthogonal wavelets. In [13], Pan *et al.* presented a hard threshold $t(j) = c\sigma_j$ for nonorthogonal wavelet expansion, where σ_j is the standard deviation of noise at the jth scale and constant $c \in [3, 4]$.

Although WT well decorrelates signals, strong intrascale and interscale dependencies between wavelet coefficients may still exist. The performance of coding and denoising would be significantly improved if such dependencies could be efficiently modeled and exploited. Liu and Moulin [27] classified the wavelet statistical models into intrascale, interscale and hybrid ones. The denoising schemes in [8], [17], [18] benefit from intrascale models. Chang *et al.* [8] proposed a spatially adaptive wavelet thresholding scheme based on context modeling. Each wavelet coefficient is modeled as a mixture of GGD with unknown slowly spatially varying parameters, and the estimation of these parameters is conditioned on a function of its neighboring coefficients. M. K. Mıhçak *et al.* [17] estimated the second-order local statistics of each coefficient with a centered square-shaped window and developed a linear minimum mean squared-error estimation (LMMSE) like denoising method. The denoising approach of Li and Orchard [18] is also LMMSE based but it models the wavelet coefficients as a mixture of edge and nonedge classes. In [26], a local contextual hidden Markov model (LCHMM) was proposed to capture the wavelet intrascale dependencies. Wavelet interscale models are also used in many other applications [12]–[15], [19], [20], [24], [25].

Manuscript received May 19, 2003; revised November 24, 2003. This work was supported in part by the Research Grant Council of the Hong Kong Special Administrative Region Grant CUHK5982/00E. This paper was recommended by Associate Editor Z. Xiong.

L. Zhang and X. Wu are with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON L8S 4L8, Canada (e-mail: johnray@mail.ece.mcmaster.ca; xwu@mail.ece.mcmaster.ca).

P. Bao is with the School of Engineering, Nanyang Technological University, Singapore 639798 (e-mail: aspbao@ntu.edu.sg).

Digital Object Identifier 10.1109/TCSVT.2005.844456

If a coefficient at a coarser scale has small magnitude, its descendants at finer scales are very likely to be small too. Shapiro [20] exploited this property and developed the well-known embedded zerotree wavelet image compression scheme. In another viewpoint, if a wavelet coefficient generated by true signal has large magnitude at a finer scale, its ascendants at coarser scales will likely be significant as well. But for those coefficients caused by noise, the magnitudes may decay rapidly along the scales. With this observation, it is expected that multiplying the wavelet coefficients at adjacent scales would strengthen the significant structures while diluting noise. Such a property has been exploited for denoising [12]–[14], step estimation [19] and edge detection [15]. The wavelet interscale dependencies have also been represented by Markov models [24], [25]. The hidden Markov models (HMMs), especially the hidden Markov tree model (HMT), proposed by Crouse [24], well characterize the joint statistics of wavelet coefficients across scales. Each coefficient is assigned with a hidden state, conditioned on which the coefficients are i.i.d. Gaussian. Some schemes adopted an interscale and intrascale hybrid model to better estimate noisy wavelet coefficients, such as Liu and Moulin [28] and Portilla *et al.* [29]. In [29], each coefficient was modeled as the product of a Gaussian random vector and a hidden multiplier variable to include adjacent scales in the conditioning local neighborhood. Liu and Moulin [27], [28] analyzed theoretically the dependency between wavelet coefficients using mutual information as a measurement. They also compared the ability of various wavelet models in encapsulating the dependency information.

The LMMSE denoising schemes in [17] and [18] exploit the wavelet intrascale dependencies. In this paper, an LMMSE-based denoising approach with an interscale model is presented by using overcomplete wavelet expansion (OWE). The optimal wavelet bases selection with respect to the proposed scheme is subsequently discussed. To exploit the wavelet intrascale dependency in our denoising approach, we spatially classify the wavelet coefficients into several clusters adaptively. With OWE, in which there is no downsampling in the decomposition, each wavelet subband has the same number of coefficients as the input image. We combine the wavelet coefficients with the same spatial location across adjacent scales as a vector, to which the LMMSE is then applied. Such an operation naturally incorporates the interscale dependencies of wavelet coefficients to improve the estimation. LMMSE is similar to soft thresholding strategy to some extent. Suppose the variable is scalar, instead of shrinking a noisy wavelet coefficient $w = x + v$ (where x is the wavelet coefficient of noiseless signal and v is that of noise) with threshold $t : \hat{x} = \mathrm{sgn}(w) \cdot \max(|w| - t, 0)$, LMMSE modifies the coefficient with a factor $c : \hat{x} = c \cdot w$, where $c = \sigma_x^2/(\sigma_x^2 + \sigma_v^2)$. σ_x^2 and σ_v^2 are the variances of signal x and noise v, respectively. Obviously, c is less than 1 so that $|\hat{x}|$ will be less than $|w|$. The energy of finally restored signal will be shrunk just like in the soft thresholding schemes.

The performance of proposed interscale LMMSE scheme is wavelet dependent. A rich library of wavelet bases have been constructed and widely used in signal processing, such as Daubechies' compactly supported orthonormal [1] and biorthogonal wavelets [2]. From denoising point of view wavelet filters should have the following two properties. One is the capability of extracting signal information from noisy wavelet coefficients. A parameter M, which is based on the mutual information of noiseless wavelet coefficients and noisy wavelet coefficients, is defined. M is proportional to the performance of the scheme. The other is that the distribution of interscale image wavelet coefficients is sufficiently close to jointly Gaussian [when the distribution is jointly Gaussian, LMMSE is equal to minimum mean square-error estimation (MMSE)]. A parameter E, which measures the difference between the Gaussian and real signal density functions, is defined and E is inversely proportional to the denoising performance. An optimal wavelet could be determined from a library of wavelets based on the M and E values of them.

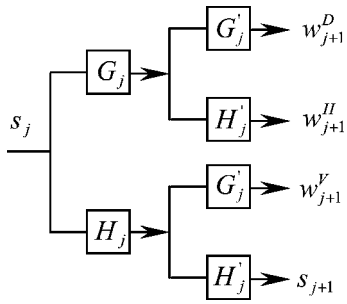

Fig. 1. One stage decomposition of the 2-D OWE. w_j^H, w_j^V and w_j^D are the wavelet coefficients at horizontal, vertical and diagonal directions.

To incorporate the intrascale dependencies into our interscale model, we classify the wavelet coefficients into several clusters adaptively by using context modeling. Context modeling gives a local discrimination of image characteristics, such as edge structures and backgrounds, according to their spatial dependencies. We extend the context modeling to interscale wavelet coefficient vector variables. The statistics of wavelet coefficients are then estimated locally from each cluster. Experiments show that context modeling improves the denoising performance.

The paper is organized as follows. In Section II, the interscale model of wavelet coefficients and the LMMSE-based denoising approach are developed. In Section III, we introduce two criteria for measuring the efficiencies of different wavelets. The optimal wavelet is selected from a library of wavelets by optimizing the tradeoff between the two criteria. Section IV improves the scheme by classifying wavelet coefficients into different clusters through context modeling. Experimental results are presented in Section V and the paper is concluded in Section VI.

II. Interscale Model and LMMSE-Based Denoising

Bi-orthogonal wavelet transform (OWT) is translation variant due to the downsampling. This will cause some visual artifacts (such as Gibbs phenomena) in threshold-based denoising [11]. It has been observed that the OWE (undecimated WT or translation-invariant WT in other names) achieves better results in noise reduction and artifacts suppression [7], [11], [13], [18]. The denoising scheme presented in this paper adopts OWE,

whose one stage two-dimensional (2-D) decomposition structure is shown in Fig. 1. No downsampling occurs but the analytic filters vary in it. Filter H_j is interpolated by putting $(2^{j-1}-1)$ zeros between each of the coefficients of original filter H_0, so does for G_j. The bandwidth decrease is accomplished by zeros padding of filters instead of downsampling of wavelet coefficients. The restored signal by OWE is an average of several circularly shifted denoised versions of the same signal by OWT, and by which the additive noise is better suppressed.

A. LMMSE of Wavelet Coefficients

Suppose the original signal f is corrupted with additive Gaussian white noise ε

$$g = f + \varepsilon \tag{1}$$

where $\varepsilon \in N(0, \sigma^2)$. Applying the OWE to the noisy signal g, at scale j yields

$$w_j = x_j + v_j \tag{2}$$

where w_j is coefficients at scale j, x_j, and v_j are the expansions of f and ε, respectively.

In this paper, the LMMSE of wavelet coefficients is employed instead of soft thresholding. Suppose the variance of v_j is σ_j^2 and that of x_j is $\sigma_{x_j}^2$. Since x_j and v_j are both zero mean, the LMMSE of x_j is

$$\hat{x}_j = c' \cdot w_j \tag{3}$$

with

$$c = \frac{\sigma_{x_j}^2}{\sigma_{x_j}^2 + \sigma_j^2}. \tag{4}$$

Since v_j is Gaussian distributed and independent of x_j, if x_j is also of Gaussian distribution, it is well known that w_j will be Gaussian and (3) is equivalent to the optimal MMSE [31]. Unfortunately, x_j obeys in general the GGD model, which reduces to Gaussian only in very special cases.

Referring to Fig. 1, term w_{j+1}^D can be written as

$$w_{j+1}^D = s_0 * L_j^D \tag{5}$$

where $*$ is the convolution operator and filter L_j^D is

$$L_j^D - H_0 * H_0' * \ldots * H_{j-1} * H_{j-1}' * G_j * G_j'. \tag{6}$$

Similarly, we have

$$w_{j+1}^H = s_0 * L_j^H, \quad w_{j+1}^V = s_0 * L_j^V \tag{7}$$

where

$$L_j^H = H_0 * H_0' * \ldots * H_{j-1} * H_{j-1}' * G_j * H_j' \tag{8}$$

$$L_j^V = H_0 * H_0' * \ldots * H_{j-1} * H_{j-1}' * H_j * G_j'. \tag{9}$$

Noise standard deviation of v_j at scale j in a direction (horizontal, vertical or diagonal) is

$$\sigma_j = \|L_{j-1}\|\sigma \tag{10}$$

where L_{j-1} is the corresponding filter (L_{j-1}^D, L_{j-1}^H or L_{j-1}^V) and $\|\bullet\|$ is the norm operator: $\|L\| = \sqrt{\sum_l \sum_k L^2(l,k)}$. The standard deviation $\sigma_{x_j}^2$ of noiseless image x_j is estimated as follows

$$\hat{\sigma}_{x_j}^2 = \sigma_{w_j}^2 - \sigma_j^2 \tag{11}$$

with

$$\sigma_{w_j}^2 = \frac{1}{M \cdot N} \sum_{m=1}^{M} \sum_{n=1}^{N} w_j^2(m,n) \tag{12}$$

where M and N are the numbers of input image rows and columns.

LMMSE is similar to soft thresholding in some sense. Notice that factor c is always less than 1, thus the magnitude of estimated wavelet coefficient $\hat{x}_j$ would be less than that of w_j. This leads to the energy shrinkage of the restored signal, same as in the soft thresholding schemes. The LMMSE-based wavelet denoising schemes proposed in [17] and [18] have achieved good results. These two methods exploited the wavelet intrascale dependencies.

B. Interscale Wavelet Model-Based LMMSE

Wavelet adjacent scales are strongly correlated and these interscale dependencies can be exploited for better signal processing results. Small magnitude coefficients at coarser scales are more likely to derive small magnitude descendents at finer scale. Contrarily, it is also found that a large magnitude wavelet coefficient produced by true signal at finer scales would yield significant coefficients at coarser scales. But the coefficients corresponding to noise decay rapidly along scales. This can be interpreted by the different singularities of signal and noise [4]. With this observation Xu *et al.* [12] multiplied the adjacent wavelet scales to sharpen the edge structures and identified significant pixels from the multiplication iteratively. Sadler and Swami [19] analyzed the multiscale products of wavelet coefficients and applied it to step detection and estimation. Zhang and Bao [15] developed an effectively edge detection approach by finding edge pixels from the scale multiplication. They also applied the wavelet scale multiplication to threshold-based denoising [14]. In [24] and [25], the HMM [24], [25] are used to represent wavelet interscale dependencies efficiently.

In this section, we apply the LMMSE-based denoising to a wavelet interscale model. It is well known that the wavelet-represented images are similar across scales, especially among the adjacent scales. In wavelet domain, the noise level decrease rapidly along scales, while signal structures are strengthened with scale increasing. So we use coarser scale information to improve finer scale estimation. Suppose the input image is decomposed into J scales. Roughly speaking, scale j is strongly correlated with scale $j+1$, but its correlations with scales $j+$

$2, j+3, \ldots, J$ will decrease rapidly. These scales would not provide much additional information to improve the estimation of scale j. Second, a significant structure has much larger local supports at coarse scales than at fine scales. At the same spatial location, the wavelet coefficients may correspond to signal at coarse scales, but to noise at fine scales. Based on these consideration, we would make no use of the measurements at the finer scale to estimate the signal at the coarser scale, and x_j is estimated only by measurements at scales j and $j+1$. We assemble the points with the same orientation at scales j and $j+1$ as a vector

$$\vec{\boldsymbol{w}}_j\,(m,n) = [w_j(m,n) \quad w_{j+1}(m,n)]^T. \tag{13}$$

Thus

$$\vec{\boldsymbol{w}}_j = \vec{\boldsymbol{x}}_j + \vec{\boldsymbol{v}}_j \tag{14}$$

with

$$\begin{aligned}\vec{\boldsymbol{x}}_j\,(m,n) &= [x_j(m,n) \quad x_{j+1}(m,n)]^T\\ \vec{\boldsymbol{v}}_j\,(m,n) &= [v_j(m,n) \quad v_{j+1}(m,n)]^T.\end{aligned} \tag{15}$$

$\vec{\boldsymbol{v}}_j$ is a Gaussian noise vector independent of $\vec{\boldsymbol{x}}_j$. The LMMSE of $\vec{\boldsymbol{x}}_j$ is then

$$\hat{\vec{\boldsymbol{x}}}_j = P_j(P_j + R_j)^{-1}\,\vec{\boldsymbol{w}}_j \tag{16}$$

where P_j and R_j are the covariance matrices of $\vec{\boldsymbol{x}}_j$ and $\vec{\boldsymbol{v}}_j$, respectively

$$\begin{aligned}P_j &= E\left[\vec{\boldsymbol{x}}_j\vec{\boldsymbol{x}}_j^{\,T}\right] = E\begin{bmatrix}x_j^2 & x_jx_{j+1}\\ x_jx_{j+1} & x_{j+1}^2\end{bmatrix}\\ R_j &= E\left[\vec{\boldsymbol{v}}_j\vec{\boldsymbol{v}}_j^{\,T}\right] = E\begin{bmatrix}v_j^2 & v_jv_{j+1}\\ v_jv_{j+1} & v_{j+1}^2\end{bmatrix}.\end{aligned} \tag{17}$$

Let us compute the components of noise covariance matrix R_j first. The diagonal element $E[v_j^2]$ is equal to σ_j^2 which can be obtained by (10). Noise variables v_j and v_{j+1} are the projections of v on different wavelet subspaces. They are correlated with correlation coefficient

$$\rho_{j,j+1} = \frac{\sqrt{\sum_l \sum_k L_{j-1}(l,k)L_j(l,k)}}{||L_{j-1}|| \cdot ||L_j||}. \tag{18}$$

v_j and v_{j+1} are jointly Gaussian and their density is

$$\begin{aligned}p(v_j, v_{j+1}) = &\frac{1}{2\pi\sigma_j\sigma_{j+1}\sqrt{1-\rho_{j,j+1}^2}}\\ &\times e^{-\frac{1}{2\left(1-\rho_{j,j+1}^2\right)}\left[\frac{v_j^2}{\sigma_j^2} - \frac{2\rho_{j,j+1}v_jv_{j+1}}{\sigma_j\sigma_{j+1}} + \frac{v_{j+1}^2}{\sigma_{j+1}^2}\right]}.\end{aligned} \tag{19}$$

Thus, the expectation $E[v_jv_{j+1}]$ is

$$E[v_jv_{j+1}] = \rho_{j,j+1}\sigma_j\sigma_{j+1}. \tag{20}$$

Each of the components of matrix P_j is estimated by

$$E[x_lx_k] \approx E[w_lw_k] - E[v_lv_k] \tag{21}$$

where $l, k = j, j+1$ and $E[w_lw_k]$ is computed as

$$E[w_lw_k] = \frac{1}{M \cdot N}\sum_{m=1}^{M}\sum_{n=1}^{N} w_l(m,n) \cdot w_k(m,n). \tag{22}$$

After the LMMSE result $\hat{\vec{\boldsymbol{x}}}_j$ is obtained, only the component $\hat{x}_j$ is extracted. Estimation of $\hat{x}_{j+1}$ would be obtained form the LMMSE result $\hat{\vec{\boldsymbol{x}}}_{j+1}$.

III. Optimal Wavelet Basis Selection

The denoising performance of the proposed LMMSE-based scheme varies with different wavelet filters. Ideally, a good wavelet filter for denoising should meet the following two requirements. One is the interscale model's ability in extracting signal information from noisy wavelet coefficients. The other is a high degree of agreement between the distribution of wavelet coefficients and Gaussian distribution. This is because the LMMSE denoising method is optimal (i.e., equivalent to optimal MMSE) only if the underlying signal distribution is Gaussian, assuming that the additive noise is Gaussian. However, for a fixed wavelet basis, the above two requirements may be in conflict with each other. In this section we develop a technique to strike a good balance between the two conflicting criteria.

A. Signal Information Extraction Criterion

For denoising purpose, it is expected that the true signal wavelet coefficients would be enhanced in the noisy environment with the interscale model. We would like to measure the signal component in the noisy coefficients for a fixed wavelet filter. As a good similarity metric, mutual information has been used in several signal processing applications [27], [30]. It computes the dependency of variables μ and ν by measuring the distance between the joint distribution $p(\mu,\nu)$, and the product of marginal distributions $p(\mu) \cdot p(\nu)$ using Kullback–Leibler measure [30]. The mutual information of μ and ν is defined as

$$I(\mu,\nu) = \sum_{\mu}\sum_{\nu} p(\mu,\nu)\log\frac{p(\mu,\nu)}{p(\mu)p(\nu)}. \tag{23}$$

The higher $I(\mu,\nu)$ is, the more information μ could provide to estimate ν or vice-versa. If μ is a function of ν, $I(\mu,\nu)$ will be infinite. Otherwise, if μ is independent with ν, obviously $I(\mu,\nu)$ is zero.

We take the mutual information of $\vec{\boldsymbol{x}}_j$ and $\vec{\boldsymbol{w}}_j$ as a measure to evaluate how much signal information could be exploited from $\vec{\boldsymbol{w}}_j$ to estimate $\vec{\boldsymbol{x}}_j$. We have derived that $\vec{\boldsymbol{v}}_j$ is Gaussian with covariance matrix R_j (refer to (17)). The covariance matrix of

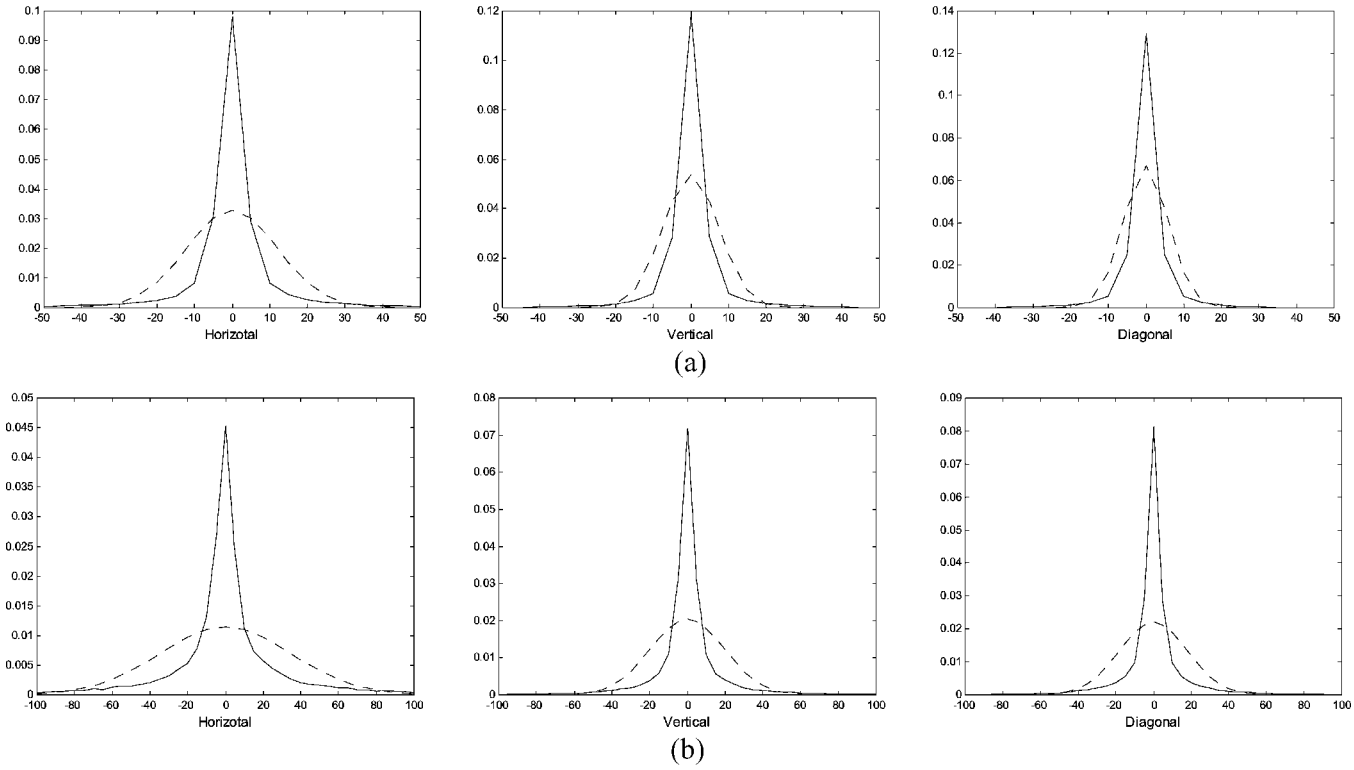

Fig. 2. Histograms (solid) of wavelet coefficients of *Lena* and the associated *Gaussian* functions (dash) with zero mean and standard deviation σ_x. (a) Scale 1; (b) Scale 2.

$\vec{\boldsymbol{x}}_j$ is P_j and we assume $\vec{\boldsymbol{x}}_j$ is also Gaussian. Since $\vec{\boldsymbol{w}}_j = \vec{\boldsymbol{x}}_j + \vec{\boldsymbol{v}}_j$, the mutual information of $\vec{\boldsymbol{x}}_j$ and $\vec{\boldsymbol{w}}_j$ is [30]

$$M_j = I(\vec{\boldsymbol{x}}_j, \vec{\boldsymbol{w}}_j) = \frac{1}{2}\log\left(\frac{|P_j + R_j|}{|R_j|}\right) \tag{24}$$

where $|\bullet|$ represents the determinant of a matrix. The criterion M_j is proportional to the performance of the proposed denoising scheme. A properly selected wavelet should yield a significant value of M_j, which means noisy coefficients $\vec{\boldsymbol{w}}_j$ could give significant information to estimate original signal $\vec{\boldsymbol{x}}_j$.

Since the image wavelet coefficients are subjected to GGD, the distribution of $\vec{\boldsymbol{x}}_j$ would be of some difference with bivariate Gaussian function. The errors so caused could be generalized into the following criterion.

B. Distribution Error Criterion

Compared with MMSE, LMMSE is suboptimal because it exploits only the second-order statistics of signal x and noise v. But it is practical and simple compared to the analytic form of MMSE which is usually impractical to implement. In the special case where x and v are zero-mean and jointly Gaussian, LMMSE will be equivalent to MMSE because the Gaussian process has only two order statistics [31]. In this paper noise v is assumed as additive Gaussian and independent of x, so that the better x follows the Gaussian distribution, the better LMMSE approximates to MMSE.

The distribution of wavelet coefficients x is often modeled as GGD [7]

$$GG_{\beta,\sigma_x}(x) = C(\beta,\sigma_x)e^{-(\alpha(\beta,\sigma_x)|x|)^B}, \quad -\infty < x < \infty, \sigma_x > 0, \beta > 0 \tag{25}$$

$$\alpha(\beta,\sigma_x) = \sigma_x^{-1}\left[\frac{\Gamma\left(\frac{3}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)}\right]^{\frac{1}{2}}$$

$$C(\beta,\sigma_x) = \frac{\beta\alpha(\beta,\sigma_x)}{2\Gamma\left(\frac{1}{\beta}\right)} \tag{26}$$

where σ_x is the standard deviation of x, β is the shape parameter and $\Gamma(t) = \int_0^\infty e^{-u}u^{t-1}du$ is the Gamma function. GGD is zero-mean and degenerates to Gaussian distribution only when $\beta = 2$. In Fig. 2 the histograms of the wavelet coefficients of image *Lena* (shown in Fig. 3(a)) at the first two scales are illustrated together with the associated Gaussian function

$$G_{\sigma_x} = \frac{1}{\sqrt{2\pi}\sigma_x}e^{-\frac{x^2}{2\sigma_x^2}}. \tag{27}$$

Obviously, there exists sharp difference between the histograms and the associated Gaussian functions. For LMMSE-based denoising, a good WT should have the histograms as close to Gaussian as possible.

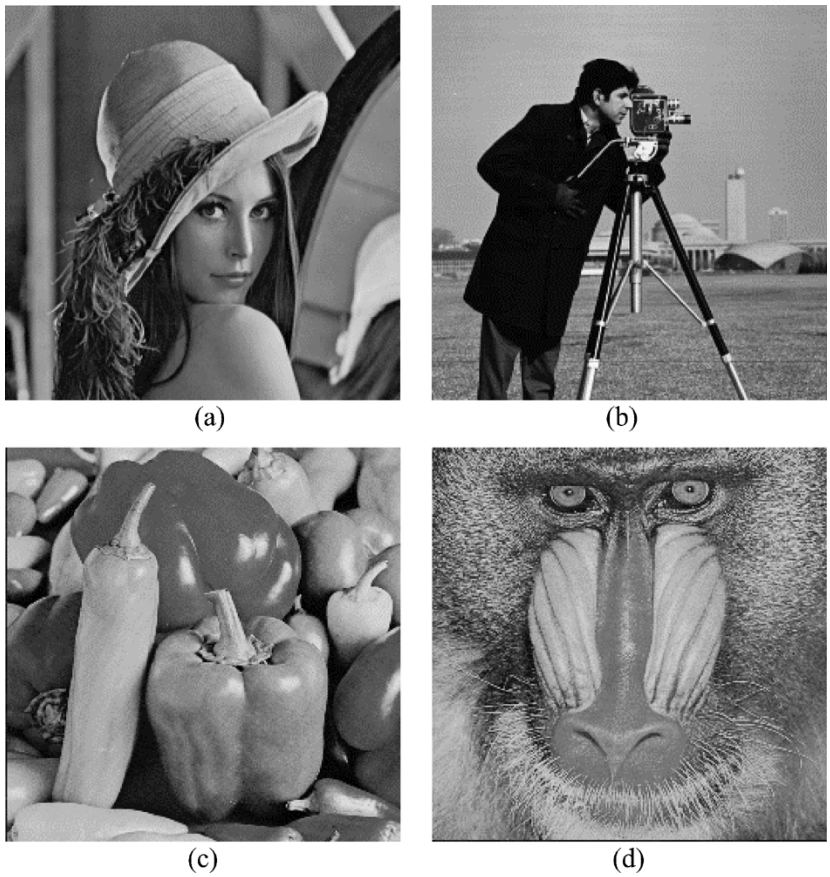

Fig. 3. Four 256 × 256 benchmark images for the experiments. (a) Lena. (b) Cameraman. (c) Peppers. (d) Baboon.

Similarly for our interscale wavelet model, it is desirable that $p(x_j, x_{j+1})$, the joint probability density function (PDF) of x_j and x_{j+1}, could be as close to jointly Gaussian as possible

$$p_g(x_j, x_{j+1}) = \frac{1}{2\pi\sigma_{x_j}\sigma_{x_{j+1}}\sqrt{1-\rho_j^2}} \times e^{-\frac{1}{2\left(1-\rho_j^2\right)}\left[\frac{x_j^2}{\sigma_{x_j}^2}-\frac{2\rho_j x_j x_{j+1}}{\sigma_{x_j}\sigma_{x_{j+1}}}+\frac{x_{j+1}^2}{\sigma_{x_{j+1}}^2}\right]} \quad (28)$$

where ρ_j is calculated as

$$\rho_j = \frac{E[x_j x_{j+1}]}{\sigma_{x_j}\sigma_{x_{j+1}}} = \frac{1}{M\cdot N\cdot\sigma_{x_j}\sigma_{x_{j+1}}} \times \sum_{m=1}^{M}\sum_{n=1}^{N} x_j(m,n)\cdot x_{j+1}(m,n). \quad (29)$$

We define the distribution error criterion as a kind of Hellinger distance

$$E_j = \sqrt{\int\int (p_g - p)^2 dx_j dx_{j+1}}. \quad (30)$$

When p and p_g are identical, the measurement E_j will reach the minimum 0. The higher the error $\widetilde{p} = p - p_g$, the higher the value of E_j, which implies that p_g worse approximates a joint Gaussian distribution, and then the LMMSE will be much inferior to the MMSE. So a good wavelet should yield a small E_j.

C. Tradeoff Parameter

As stated above, the denoising performance increases in M_j but decreases in E_j. Therefore, a good wavelet basis for denoising should aim at maximizing M_j and minimizing E_j, which are in general conflicting criteria. In order to balance the criteria we introduce a tradeoff parameter r_j. Intuitively, one may want to set $r_j = M_j/E_j$. However, the metric units of M_j and E_j differ by a logarithmic factor. Namely, M_j is a weighted sum of logarithmic functions of the PDF of wavelet coefficients, while E_j is a direct difference function of between two PDFs. Accordingly we adjust the scale of E_j such that $E_j = \log A$, and define

$$r_j = \frac{M_j}{A} = M_j \cdot e^{-E_j} \quad (31)$$

where e^{-E_j} ranges from 0 to 1, and when error E_j runs to zero it reaches the maximum value 1. The optimal wavelet can be selected from a library of wavelets by maximizing r_j. In this paper, we focus on the widely used compactly supported orthogonal and biorthogonal wavelets constructed by Daubechies *et al.* [1], [2].

We denote Daubechies' orthogonal wavelets [1] by $Dau(N)$, where $N = 1, 2, \ldots, \infty$ is the vanishing moment of the wavelet whose filter length will be $2N$. The biorthogonal wavelet in [2] is denoted by $CDF(N, N')$, where N is the vanishing moment

TABLE I
VALUES OF M_j, E_j AND r_j FOR IMAGE *LENA*

Lena			*CDF*(1,1)	*CDF*(1,3)	*CDF*(2,2)	*CDF*(2,4)	*CDF*(3,3)	*Dau*(2)	*Dau*(3)	*Dau*(4)
M_j	j=1	*H*	1.1357	**1.2148**	0.6473	0.7087	0.4718	1.0087	0.9553	0.9293
		V	0.6221	**0.6867**	0.3104	0.3401	0.2375	0.4761	0.4291	0.4069
		D	0.3711	**0.4134**	0.2406	0.2634	0.1985	0.3438	0.3425	0.3436
	j=2	*H*	2.7660	**2.9806**	1.9688	2.1933	1.5008	2.7963	2.7895	2.7797
		V	1.8134	**1.9860**	1.1129	1.2613	0.8741	1.6140	1.5484	1.5098
		D	1.3403	**1.5179**	0.9599	1.1037	0.8279	1.3425	1.3526	1.3595
E_j	j=1	*H*	0.1387	0.1300	0.1563	0.1518	**0.1297**	0.1549	0.1546	0.1529
		V	0.1853	0.1772	0.1993	0.1971	**0.1585**	0.2191	0.2241	0.2228
		D	0.2576	**0.2422**	0.2778	0.2728	0.2444	0.2599	0.2544	0.2503
	j=2	*H*	0.0606	0.0508	0.0521	0.0487	**0.0323**	0.0594	0.0544	0.0511
		V	0.0956	0.0794	0.0823	0.0794	**0.0490**	0.1038	0.1040	0.1010
		D	0.1462	0.1235	0.1114	0.1058	**0.0698**	0.1340	0.1262	0.1206
r_j	j=1	*H*	0.9886	**1.0667**	0.5536	0.6089	0.4144	0.8639	0.8184	0.7976
		V	0.5169	**0.5752**	0.2543	0.2793	0.2027	0.3824	0.3429	0.3256
		D	0.2868	**0.3245**	0.1822	0.2005	0.1555	0.2651	0.2655	0.2675
	j=2	*H*	2.6034	**2.8331**	1.8689	2.0890	1.4531	2.6350	2.6419	2.6413
		V	1.6481	**1.8344**	1.0249	1.1651	0.8322	1.4549	1.3955	1.3647
		D	1.1580	**1.3416**	0.8587	0.9930	0.7721	1.1742	1.1922	1.2051

TABLE II
VALUES OF M_j, E_j AND r_j FOR IMAGE *CAMERAMAN*

Cameraman			*CDF*(1,1)	*CDF*(1,3)	*CDF*(2,2)	*CDF*(2,4)	*CDF*(3,3)	*Dau*(2)	*Dau*(3)	*Dau*(4)
M_j	j=1	*H*	1.2958	**1.4036**	0.9283	1.0053	0.7317	1.2837	1.2685	1.2590
		V	1.1325	**1.1928**	0.7714	0.8207	0.5565	1.1019	0.9789	0.8930
		D	0.4457	**0.4830**	0.2962	0.3194	0.2585	0.3918	0.3826	0.3801
	j=2	*H*	2.8121	**3.0672**	2.2288	2.4723	1.8128	2.9072	2.9226	2.9197
		V	2.4912	**2.6447**	1.7697	1.9338	1.3284	2.4337	2.2705	2.2335
		D	1.3869	**1.5432**	0.9590	1.0885	0.8251	1.3207	1.3032	1.2971
E_j	j=1	*H*	0.3353	0.3048	0.2898	0.2793	**0.2104**	0.3335	0.3170	0.3060
		V	0.2608	0.2366	0.2304	0.2222	**0.1840**	0.2406	0.2277	0.2165
		D	0.3997	0.3742	0.3660	0.3596	**0.3092**	0.3645	0.3511	0.3397
	j=2	*H*	0.1839	0.1557	0.1498	0.1426	**0.0809**	0.1961	0.1814	0.1664
		V	0.1589	0.1244	0.1048	0.0981	**0.0605**	0.1310	0.1151	0.1056
		D	0.2962	0.2520	0.1899	0.1837	**0.1115**	0.2514	0.2325	0.2160
r_j	j=1	*H*	0.9267	**1.0349**	0.6948	0.7604	0.5929	0.9196	0.9239	0.9271
		V	0.8725	**0.9416**	0.6127	0.6572	0.4629	0.8663	0.7796	0.7191
		D	0.2988	**0.3322**	0.2054	0.2230	0.1898	0.2721	0.2693	0.2706
	j=2	*H*	2.3397	**2.6198**	1.9187	2.1438	1.6719	2.3894	2.4378	2.4703
		V	2.1253	**2.3354**	1.5936	1.7532	1.2505	2.1349	2.0235	2.0097
		D	1.0314	**1.1994**	0.7931	0.9058	0.7381	1.0271	1.0328	1.0451

of analytic wavelet and N' is that of synthetic wavelet. The set of $Dau(N)$ wavelets (except for $Dau(1)$, i.e., $CDF(1,1)$ or Haar wavelet) are lack of (anti)symmetry, which is an important property in signal and image processing. Biorthogonal wavelets $CDF(N,N')$ trade the orthogonality for the (anti-)symmetry.

We use four 256 × 256 benchmark images *Lena*, *Cameraman*, *Peppers*, and *Baboon*, shown in Fig. 3, to compute their M_j, E_j and r_j values with respect to eight wavelets: $CDF(1,1), CDF(1,3), CDF(2,2), CDF(2,4), CDF(3,3)$, $Dau(2)$, $Dau(3)$, and $Dau(4)$. In calculating M_j, we set the noise level σ as 25 (other noise level values would reach the similar analysis results). The PDF function $p(x_j, x_{j+1})$ for calculating E_j is taken as the histogram of image wavelet coefficients. The noise decreases rapidly along wavelet scales and most of its energy is concentrated at the first three scales. In wavelet based denoising, three-scale decomposition is well accepted because noise is greatly smoothed in the third-level low frequency band. Decomposing an image into more than three scales would not yield much additional improvement in noise reduction. In Tables I–IV, we list the values of M_j, E_j and r_j when $j = 1$ and $j = 2$. These results represent the information of the first three wavelet scales. (The letters *H*, *V* and *D* in Tables I–IV indicate the horizontal, vertical and diagonal subbands, respectively.)

The best values of M_j, E_j, and r_j are highlighted in Tables I–IV. From the experimental results, it can be observed that $CDF(1,3)$ is obviously the best of the eight wavelets. Its r_j values are almost always higher than that of other wavelets. $CDF(1,3)$ (i.e., $Dau(1)$) and $Dau(2)$ are also proper selections. Biorthogonal wavelets $CDF(2,2)$, $CDF(2,4)$, $CDF(3,3)$ are inferior. They are not suitable for the proposed denoising scheme. The experiments in Section V validated these observations.

TABLE III
VALUES OF M_j, E_j AND r_j FOR IMAGE *PEPPERS*

Peppers			*CDF*(1,1)	*CDF*(1,3)	*CDF*(2,2)	*CDF*(2,4)	*CDF*(3,3)	*Dau*(2)	*Dau*(3)	*Dau*(4)
M_j	j=1	*H*	1.0977	**1.1627**	0.5987	0.6517	0.4004	0.9858	0.8785	0.8330
		V	0.8441	**0.9112**	0.4416	0.4855	0.3204	0.7000	0.6510	0.6309
		D	0.3130	**0.3406**	0.1757	0.1882	0.1416	0.2604	0.2471	0.2414
	j=2	*H*	2.6914	**2.8872**	1.8838	2.0997	1.3996	2.7163	2.6581	2.6542
		V	2.2488	**2.4282**	1.4491	1.6354	1.0899	2.1452	2.1014	2.0758
		D	1.2444	**1.4004**	0.8082	0.9402	0.6439	1.2012	1.2033	1.2016
E_j	j=1	*H*	0.1260	0.1159	0.1247	0.1202	**0.0956**	0.1307	0.1257	0.1212
		V	0.1294	0.1233	0.1240	0.1203	**0.0879**	0.1408	0.1382	0.1345
		D	0.1652	0.1545	0.1652	0.1625	0.1445	0.1528	0.1435	**0.1394**
	j=2	*H*	0.0564	0.0464	0.0461	0.0426	**0.0262**	0.0513	0.0449	0.0415
		V	0.0640	0.0538	0.0525	0.0501	**0.0302**	0.0667	0.0656	0.0623
		D	0.0919	0.0768	0.0605	0.0573	**0.0365**	0.0774	0.0690	0.0648
r_j	j=1	*H*	0.9678	**1.0355**	0.5285	0.5779	0.3639	0.8650	0.7747	0.7379
		V	0.7417	**0.8055**	0.3901	0.4305	0.2934	0.6080	0.5670	0.5515
		D	0.2653	**0.2919**	0.1490	0.1600	0.1226	0.2235	0.2141	0.2100
	j=2	*H*	2.5437	**2.7563**	1.7990	2.0121	1.3633	2.5804	2.5414	2.5463
		V	2.1093	**2.3009**	1.3750	1.5555	1.0575	2.0068	1.9681	1.9504
		D	1.1351	**1.2969**	0.7608	0.8879	0.6208	1.1117	1.1231	1.1262

TABLE IV
VALUES OF M_j, E_j AND r_j FOR IMAGE *BABOON*

Baboon			*CDF*(1,1)	*CDF*(1,3)	*CDF*(2,2)	*CDF*(2,4)	*CDF*(3,3)	*Dau*(2)	*Dau*(3)	*Dau*(4)
M_j	j=1	*H*	1.1828	**1.2255**	0.9781	1.0109	0.9001	1.1410	1.1333	1.1300
		V	1.4842	1.5654	1.3382	1.3834	1.2301	1.5108	1.5858	**1.5961**
		D	0.9564	**0.9837**	0.8762	0.8914	0.8386	0.9450	0.9369	0.9332
	j=2	*H*	2.1476	**2.2836**	1.7009	1.8440	1.4895	2.1443	2.1624	2.1780
		V	2.2889	2.4305	1.9060	2.0370	1.6372	2.3434	2.4849	**2.4945**
		D	1.4665	**1.5724**	1.2928	1.3673	1.2161	1.4835	1.4922	1.4961
E_j	j=1	*H*	0.0327	0.0302	0.0287	0.0276	**0.0230**	0.0303	0.0292	0.0284
		V	0.0435	0.0423	0.0392	0.0387	**0.0292**	0.0457	0.0460	0.0451
		D	0.0502	0.0469	0.0494	0.0484	0.0464	0.0465	0.0457	**0.0452**
	j=2	*H*	0.0173	0.0153	0.0110	0.0108	**0.0080**	0.0153	0.0148	0.0138
		V	0.0260	0.0240	0.0192	0.0194	**0.0113**	0.0282	0.0291	0.0282
		D	0.0275	0.0240	0.0188	0.0182	**0.0146**	0.0245	0.0228	0.0224
r_j	j=1	*H*	1.1447	**1.1890**	0.9504	0.9834	0.8796	1.1069	1.1007	1.0983
		V	1.4211	1.5006	1.2867	1.3309	1.1947	1.4432	1.5146	**1.5257**
		D	0.9096	**0.9387**	0.8340	0.8493	0.8006	0.9020	0.8950	0.8919
	j=2	*H*	2.1108	**2.2490**	1.6823	1.8242	1.4777	2.1118	2.1306	2.1482
		V	2.2302	2.3728	1.8699	1.9978	1.6188	2.2782	2.4137	**2.4251**
		D	1.4267	**1.5352**	1.2687	1.3425	1.1984	1.4476	1.4586	1.4630

IV. CLUSTERING THE WAVELET COEFFICIENTS BY CONTEXT MODELING

As seen in Fig. 2, the histograms of wavelet coefficients are not very close to the *Gaussian* distributions. In fact, it is not appropriate to model all the wavelet coefficients in one subband with only one random variable. For example, the edge pixels are of large magnitude while the background pixels being of small magnitude. They have very different variances and it would introduce numerous errors in PDF if considering them as the samples of the same Gaussian variable. To estimate the statistics of wavelet coefficients more accurately and adaptively, we model them with several variables. The context modeling technique [8], [21]–[23], which was widely used in differentiating and gathering pixels with some similarities but not necessarily spatially adjacent, is a good technique for the classification. The statistics could then be estimated locally within each cluster of pixels. Chang *et al.* [8] presented a similar work with their *BayesShrink* denoising scheme. By computing the context of each wavelet coefficient, Chang *et al.* estimated the standard deviation of it with a collection of pixels whose context values fall into a specified field.

The context value of a given coefficient is defined as a function of its neighbors. The weighted average of its adjacent pixels is often employed. In [8], $C_j(m,n)$, the context value of noisy wavelet coefficient $w_j(m,n)$ is calculated as the weighted average of the magnitude of its neighbors

$$C_j(m,n) = \boldsymbol{u}_j^{m,n}\boldsymbol{h}_j \tag{32}$$

where $\boldsymbol{u}_j^{m,n}$ is a 1×9 vector whose elements are the absolute value of $w_j(m,n)$'s eight nearest neighbors plus its parent at

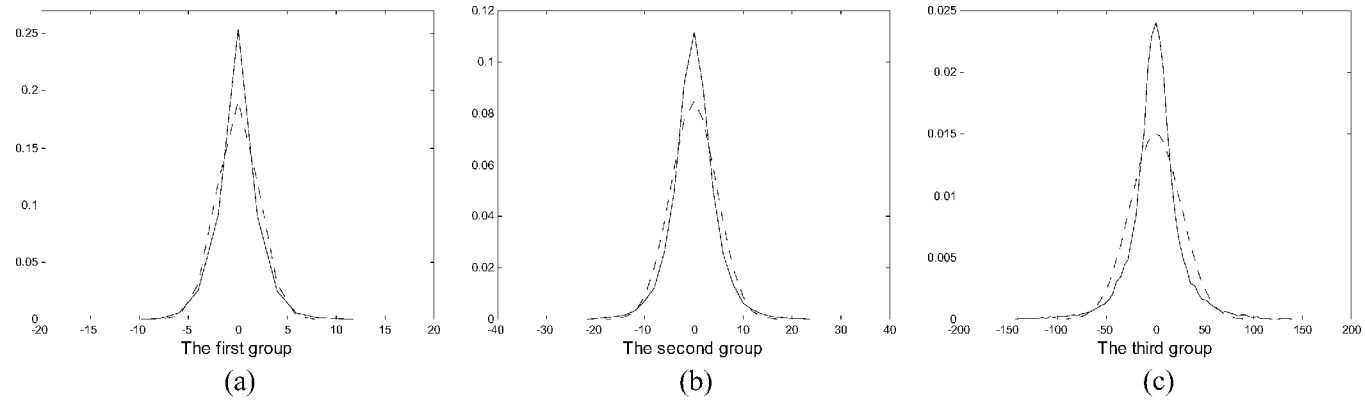

Fig. 4. Clustering the diagonal wavelet coefficients of *Lenna* at the second scale into three clusters by the context values. Solid curves are the histograms of each cluster, and dash curves are the associated *Gaussian* functions with zero mean and standard deviation σ_x of the cluster. Compared with the curves in Fig. 2, obviously the two curves are more closed. (a) The first cluster. (b) The second cluster. (c) The third cluster.

scale $j+1$. Weight $\boldsymbol{h}_j$ is a 9×1 vector. The weight $\boldsymbol{h}_j$ is determined by the least square estimation

$$\boldsymbol{h}_j \left(\boldsymbol{U}_j^T \boldsymbol{U}_j\right)^{-1} \boldsymbol{U}_j^T |\boldsymbol{Y}_j| \tag{33}$$

where $\boldsymbol{U}_j$ is a $M \cdot N \times 9$ matrix with each row being $\boldsymbol{u}_j^{m,n}$ and $\boldsymbol{Y}_j$ is an $M \cdot N \times 1$ vector containing all elements of $w_j(m,n)$.

With context modeling, the coefficients of similar natures can be well clustered. By sorting the context values $C_j(m,n)$ in the ascending order, the wavelet coefficients $w_j(m,n)$ could be classified into several clusters, in which the coefficients are assumed having the same statistics. Suppose each cluster has L point and denote by Θ_j^i the i^{th} cluster, the variance of wavelet coefficients in cluster Θ_j^i is estimated by

$$\hat{\sigma}_{x_j}^2 = \frac{1}{L} \sum_{w_j(m,n) \in \Theta_j^i} w_j^2(m,n) - \sigma_j^2. \tag{34}$$

Take the diagonal wavelet coefficients at the second scale of *Lena* as an example, we calculate their context values and then divide them into 3 clusters. In Fig. 4, the histogram of each cluster is illustrated comparing with the corresponding Gaussian function $G_{\sigma_{X_j}}(x)$. It can be seen that the two curves are much closer than those in Fig. 2(b), in which the whole band is modeled with only one random variable. With the context-based clustering, the statistical estimation of the coefficient is more accurate and adaptive.

For our interscale model, calculation of the context value of vector $\vec{\boldsymbol{w}}_j$ is straightforward. Denote by $\vec{\boldsymbol{w}}_{j:(m,n)}\ (i)$, $i = 1, \ldots, 8$ the absolute value of the 8 neighboring elements of $\vec{\boldsymbol{w}}_j\ (m,n)$. And let

$$\vec{\boldsymbol{u}}_j^{m,n} = \left[\vec{\boldsymbol{w}}_{j:(m,n)}^T\ (1) \cdots \vec{\boldsymbol{w}}_{j:(m,n)}^T\ (8)\right]. \tag{35}$$

The context value of $\vec{\boldsymbol{w}}_j\ (m,n)$ is defined as

$$\vec{\boldsymbol{C}}_j\ (m,n) = \vec{\boldsymbol{u}}_j^{m,n} \vec{\boldsymbol{h}}_j \tag{36}$$

where $\vec{\boldsymbol{h}}_j$ is a 16×2 weighted matrix and it is calculated by the least square estimate

$$\vec{\boldsymbol{h}}_j = \left(\vec{\boldsymbol{U}}_j^T \vec{\boldsymbol{U}}_j\right)^{-1} \vec{\boldsymbol{U}}_j^T\ |\ \vec{\boldsymbol{Y}}\ | \tag{37}$$

$\vec{\boldsymbol{U}}_j$ is a $M \cdot N \times 16$ matrix with each row being $\vec{\boldsymbol{u}}_j^{m,n}$ and $\vec{\boldsymbol{Y}}$ is a $M \cdot N \times 2$ matrix with each row being $\vec{\boldsymbol{w}}_j^T\ (m,n)$.

$\vec{\boldsymbol{C}}_j\ (m,n)$ is a 1×2 vector. We cluster $\vec{\boldsymbol{w}}_j\ (m,n)$ according to the position of $\vec{\boldsymbol{C}}_j\ (m,n)$ in the 2-D Euclidean space. Let the x-coordinate represent the first element of $\vec{\boldsymbol{C}}_j\ (m,n)$ and y-coordinate represent another. We first divide the plane into several regions by evenly splitting x-coordinate with a preset step length (which is $2\sigma_j$ in our experiments), and then evenly split each region by y-coordinate. The so divided regions may have different numbers of context data $\vec{\boldsymbol{C}}_j\ (m,n)$. Each vector $\vec{\boldsymbol{w}}_j\ (m,n)$ that belongs to the ith cluster Θ_j^i is assumed to possess the same the covariance matrix, which is estimated from all the data in Θ_j^i. The LMMSE scheme described in Section II-B is then applied to each cluster Θ_j^i.

V. Experiments

This section compares the proposed scheme with other popular denoising schemes. The four benchmark images in Fig. 3 are used for the experiments. As stated in Section III, the values of r_j imply that wavelet $CDF(1,3)$ would be most suitable for the proposed scheme. Other four wavelets $CDF(1,1)$, $CDF(2,2)$, $Dau(2)$, and $Dau(4)$ are employed for comparison. The noisy images are simulated by adding Gaussian white noise $\varepsilon \sim N(0, \sigma^2)$ on the original images.

In threshold-based (hard or soft) denoising schemes, the wavelet coefficients whose magnitudes are below a threshold will be set to 0. The corresponding pixels are generally noise predominated and thus the thresholding of these coefficients is safely a structure-preserving denoising process. We apply the LMMSE only to those coefficients above a threshold and shrink those below the threshold to 0. Here the threshold applied to w_j is set as $t_j = 2.5\sigma_j$.

TABLE V
PSNR (dB) RESULTS OF THE DENOISING SCHEMES FOR *LENA* WITH DIFFERENT NOISE LEVELS

Lena		*Dau*(4)	*Dau*(2)	*CDF*(1,1)	*CDF*(1,3)	*CDF*(2,2)
$\sigma = 20$	*SCH*1	29.27	29.33	29.19	--	--
	*SCH*2	29.35	29.30	28.97	28.99	29.16
	*SCH*3	29.29	29.32	29.14	29.05	28.81
	*SCH*4	28.64	28.84	29.01	29.23	27.86
	*SCH*5	29.10	29.28	29.46	**29.55**	28.74
$\sigma = 25$	*SCH*1	28.04	28.12	28.09	--	--
	*SCH*2	28.10	28.08	27.79	27.80	27.90
	*SCH*3	28.14	28.15	28.01	28.00	27.61
	*SCH*4	27.45	27.70	27.92	28.16	26.77
	*SCH*5	27.98	28.20	28.35	**28.42**	27.60
$\sigma = 30$	*SCH*1	27.10	27.20	27.19	--	--
	*SCH*2	27.11	27.15	26.90	26.87	26.96
	*SCH*3	27.22	27.25	27.15	27.24	26.69
	*SCH*4	26.70	26.89	27.06	27.30	26.11
	*SCH*5	27.08	27.30	27.54	**27.61**	26.68

TABLE VI
PSNR (dB) RESULTS OF THE DENOISING SCHEMES FOR *CAMERAMAN* WITH DIFFERENT NOISE LEVELS

Cameraman		*Dau*(4)	*Dau*(2)	*CDF*(1,1)	*CDF*(1,3)	*CDF*(2,2)
$\sigma = 20$	*SCH*1	28.02	28.04	28.07	--	--
	*SCH*2	28.43	28.46	28.49	28.32	28.37
	*SCH*3	28.42	28.63	28.79	28.87	28.25
	*SCH*4	27.39	27.83	28.28	28.12	26.88
	*SCH*5	28.32	28.71	29.12	**29.16**	28.20
$\sigma = 25$	*SCH*1	26.82	27.08	27.21	--	--
	*SCH*2	27.22	27.28	27.35	27.16	27.13
	*SCH*3	27.28	27.52	27.73	27.84	26.96
	*SCH*4	26.21	26.54	27.04	27.06	25.36
	*SCH*5	27.19	27.60	28.05	**28.07**	27.04
$\sigma = 30$	*SCH*1	25.83	26.02	26.21	--	--
	*SCH*2	26.20	26.25	26.36	26.18	26.10
	*SCH*3	26.34	26.49	26.69	26.74	25.89
	*SCH*4	25.30	25.71	26.10	26.15	24.84
	*SCH*5	26.30	26.70	27.18	**27.20**	26.06

TABLE VII
PSNR (dB) RESULTS OF THE DENOISING SCHEMES FOR *PEPPERS* WITH DIFFERENT NOISE LEVELS

Peppers		*Dau*(4)	*Dau*(2)	*CDF*(1,1)	*CDF*(1,3)	*CDF*(2,2)
$\sigma = 20$	*SCH*1	29.10	29.25	29.19	--	--
	*SCH*2	28.90	28.96	28.75	28.69	28.80
	*SCH*3	29.21	29.33	29.21	29.30	28.83
	*SCH*4	28.47	28.86	29.22	29.49	27.80
	*SCH*5	29.05	29.42	29.76	**29.82**	28.83
$\sigma = 25$	*SCH*1	27.81	28.04	28.20	--	--
	*SCH*2	27.73	27.82	27.66	27.62	27.60
	*SCH*3	28.02	28.11	28.21	28.26	27.57
	*SCH*4	27.31	27.73	28.12	28.41	26.64
	*SCH*5	27.97	28.34	28.71	**28.78**	27.62
$\sigma = 30$	*SCH*1	26.78	26.98	27.22	--	--
	*SCH*2	26.73	26.82	26.70	26.63	26.54
	*SCH*3	27.02	27.16	27.15	27.23	26.49
	*SCH*4	26.48	26.84	27.19	27.52	25.77
	*SCH*5	27.01	27.40	27.82	**27.90**	26.60

TABLE VIII
PSNR (dB) RESULTS OF THE DENOISING SCHEMES FOR *BABOON* WITH DIFFERENT NOISE LEVELS

Baboon		*Dau*(4)	*Dau*(2)	*CDF*(1,1)	*CDF*(1,3)	*CDF*(2,2)
$\sigma = 20$	*SCH*1	24.84	24.14	24.71	--	--
	*SCH*2	25.17	25.19	25.16	25.17	25.13
	*SCH*3	**24.75**	24.52	24.69	24.73	24.60
	*SCH*4	23.21	23.28	23.30	23.39	22.88
	*SCH*5	23.93	23.90	23.97	24.05	23.69
$\sigma = 25$	*SCH*1	23.65	23.64	23.63	--	--
	*SCH*2	23.94	**23.96**	23.93	23.92	23.88
	*SCH*3	23.48	23.48	22.47	23.50	23.33
	*SCH*4	22.05	22.11	22.12	22.21	21.72
	*SCH*5	22.74	22.72	22.73	22.79	22.54
$\sigma = 30$	*SCH*1	22.63	22.64	22.65	--	--
	*SCH*2	**23.07**	23.03	23.05	23.02	22.98
	*SCH*3	22.58	22.57	22.56	22.59	22.44
	*SCH*4	21.32	21.31	21.35	21.44	21.12
	*SCH*5	21.91	21.90	21.89	22.01	21.76

The denoising schemes in [8], [17], [18] are used for comparison with the proposed scheme. It should be noted that the images used here are 256 × 256, while the images used in [8], [17], and [18] are 512 × 512. At the same noise level, the denoising results of high resolution images are much better than those of low resolution images. For convenience, we denote the spatially adaptive *BayesShrink* of Chang *et al.* [8] by $SCH1$, the locally adaptive LMMSE-based scheme of Mihçak *et al.* [17] by $SCH2$ and the LMMSE-based scheme of Li *et al.* [18], which models the wavelet coefficients as nonedge and edge groups, by $SCH3$. All schemes are implemented with the OWE of the five orthogonal and biorthogonal wavelets $CDF(1,3)$, $CDF(1,1)$, $CDF(2,2)$, $Dau(2)$, and $Dau(4)$ for comparison fairness. The proposed scheme without context modeling is denoted by $SCH4$ and the counterpart with context modeling is denoted by $SCH5$.

Tables V–VIII list the peak signal-to-noise ratio (PSNR) results of the five schemes on the benchmark images in Fig. 3 corrupted by different levels of additive Gaussian noise. From the first three tables we see that the context modeling would improve the denoising performance and the scheme $SCH5$ outperforms the other four schemes. For comparison of wavelet filters, obviously $CDF(1,3)$ is the best one, whose denoising results by SCH5 were highlighted in Tables V–VII. $CDF(1,1)$ and $Dau(2)$ also have better performances. The wavelet $CDF(2,2)$ yields the worst results. Notice that these conclusions are in accordance with the wavelet filter analyzes in Section III.

In Fig. 5, we illustrated a set of denoising results of image *Lena*. Fig. 5(a) is the noisy *Lena* where the noise level is $\sigma = 20$. Fig. 5(b)–(d) are the denoised versions by $SCH1$, $SCH2$, and $SCH3$ with wavelet $Dau(2)$. Fig. 5(e), (f) are the denoised images by $SCH4$ and $SCH5$ with wavelet $CDF(1,3)$. It is observed that $SCH1$ and $SCH4$ over-smooth the image a little, and in $SCH5$ the edge structures are well preserved while reducing noise.

Although the proposed scheme works well for images *Lena*, *Cameraman*, and *Peppers*, it does not give satisfying results for *Baboon* (referring to Table VIII). $SCH5$ works worse than the three schemes $SCH1 \sim SCH3$. This is because image *Baboon* has many fine "*hair*" structures. These structures are weakly correlated and similar to white noise to some extent. They possess little interscale dependencies in wavelet domain. For quantitative measurement, let us calculate the mutual information of wavelet coefficients x_j and x_{j+1} at adjacent scales. Referring to (23) and let

$$I_j^x = I(x_j, x_{j+1}) \tag{38}$$

Fig. 5. Denoising results of Lena. (a) Noisy Lena ($\sigma = 20$). (b), (c), and (e) Denoised images by $SCH1$, $SCH2$, and $SCH3$ with wavelet $Dau(4)$. (e)–(f) Denoised images by $SCH4$ and $SCH5$ with wavelet $CDF(1,3)$.

be the mutual information between x_j and x_{j+1}. In Table IX, the values of I_j^x when $j = 1, 2$ are listed for the four test images. It can be noticed that the I_j^x values for *Baboon* are much smaller than those for *Lena*, *Cameraman*, and *Peppers*. It implies that not much information would be conveyed from scale $j+1$ to scale j for updating the estimation of x_j. So the proposed denoising scheme, mainly based on an interscale wavelet model in exploiting interscale dependency information, would not present its merits for images such as *Baboon*.

VI. Conclusion

In this paper, we presented an LMMSE-based denoising scheme with a wavelet interscale model and discussed the optimal wavelet basis selection for it. With OWE the wavelet coefficients at the same spatial locations at two adjacent scales are represented as a vector and the LMMSE is applied to the vector. The wavelet interscale dependencies are thus exploited to improve the signal estimation. The performance of the scheme is wavelet filters dependent. We proposed two criteria to determine the optimal wavelet for the scheme. One is to measure the signal information encapsulating ability from noisy environment. This criterion is proportional to the denoising efficiency. The other is to measure the wavelet coefficients distribution difference with joint Gaussian function and this criterion is inversely proportional to denoising performance. The optimal wavelet could be determined by optimizing the tradeoff of the two criteria from a library of wavelets. In this

TABLE IX
MUTUAL INFORMATION I_j^x OF ADJACENT TWO WAVELET SCALES FOR THE FOUR TEST IMAGES

I_j^x			*CDF*(1,1)	*CDF*(1,3)	*CDF*(2,2)	*CDF*(2,4)	*CDF*(3,3)	*Dau*(2)	*Dau*(3)	*Dau*(4)
Lena	j=1	*H*	0.9816	0.9071	0.5140	0.4521	0.2593	0.5254	0.3509	0.2608
		V	0.7108	0.6744	0.4104	0.3517	0.2207	0.4148	0.2939	0.2288
		D	0.4637	0.3933	0.2222	0.2008	0.1613	0.2534	0.1796	0.1780
	j=2	*H*	1.1147	0.9296	0.5463	0.4899	0.3075	0.6364	0.4391	0.3627
		V	0.8949	0.7880	0.5014	0.4428	0.2946	0.5460	0.4059	0.3257
		D	0.6207	0.5504	0.3736	0.3575	0.3180	0.3869	0.2878	0.2824
Cameraman	j=1	*H*	0.5693	0.4935	0.4561	0.4033	0.2912	0.4858	0.3982	0.3568
		V	0.6830	0.6117	0.4673	0.4077	0.3020	0.4969	0.3770	0.3225
		D	0.4634	0.3854	0.2684	0.2424	0.1944	0.2973	0.2303	0.2232
	j=2	*H*	0.7347	0.6559	0.5679	0.5346	0.4172	0.6276	0.5843	0.5207
		V	0.8843	0.7456	0.5861	0.5318	0.4042	0.6416	0.5265	0.4464
		D	0.6226	0.5827	0.4443	0.4278	0.3741	0.4593	0.3714	0.3637
Peppers	j=1	*H*	0.7612	0.7094	0.3992	0.3428	0.1708	0.4152	0.2752	0.2063
		V	0.6852	0.6319	0.3420	0.2894	0.1563	0.3692	0.2420	0.1907
		D	0.3398	0.2740	0.1581	0.1369	0.1149	0.1743	0.1041	0.0993
	j=2	*H*	1.0660	0.8645	0.5316	0.4622	0.2695	0.6052	0.4348	0.3487
		V	0.9615	0.8113	0.4461	0.3990	0.2470	0.5400	0.3936	0.0018
		D	0.5680	0.5047	0.2704	0.2572	0.2160	0.3059	0.1949	0.1889
Baboon	j=1	*H*	0.3736	0.3727	0.2814	0.2156	0.1268	0.2862	0.1815	0.1280
		V	0.3815	0.3387	0.2926	0.2350	0.1579	0.3069	0.2241	0.1784
		D	0.3356	0.2391	0.1813	0.1529	0.1308	0.1949	0.1416	0.1371
	j=2	*H*	0.5461	0.5226	0.3277	0.2769	0.1988	0.3737	0.2618	0.1924
		V	0.5049	0.4694	0.3267	0.2842	0.2278	0.3782	0.2881	0.2444
		D	0.3398	0.2690	0.1960	0.1733	0.2142	0.1994	0.1390	0.1311

paper, we considered eight typical wavelets and observed that biorthogonal $CDF(1,3)$ would give the best performance. This observation was validated by the experiments. Finally, context modeling techniques were employed to cluster wavelet coefficients. The adaptively spatial classification of wavelet pixels reduces the statistics estimation error and subsequently improves the denoising performance.

Although the proposed scheme outperforms other popular denoising schemes for most of the images, it may not be a suitable method for images that are weakly correlated in scale spaces (for example, *Baboon* image). For such images the wavelet interscale dependency is typically very low, and the proposed model would be unable to take the advantage of interscale dependencies to yield reasonable gain for denoising.

REFERENCES

[1] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Comm. Pure Appl. Math.*, vol. 41, pp. 909–996, 1988.

[2] A. Cohen, I. Daubechies, and J. C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Comm. Pure Appl. Math.*, vol. 45, pp. 485–560, 1992.

[3] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[4] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, no. 7, pp. 710–732, Jul. 1992.

[5] M. Vetterli and C. Herley, "Wavelet and filter banks: theory and design," *IEEE Trans. Signal Process.*, vol. 40, no. 9, pp. 2207–2232, Sep. 1992.

[6] R. W. Dijkerman and R. R. Mazumdar, "Wavelet representations of stochastic processes and multiresolution stochastic models," *IEEE Trans. Signal Process.*, vol. 42, no. 7, pp. 1640–1652, Jul. 1994.

[7] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1532–1546, Sep. 2000.

[8] ——, "Spatially adaptive wavelet thresholding with context modeling for image denoising," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1522–1531, Sep. 2000.

[9] D. L. Donoho, "De-noising by soft thresholding," *IEEE Trans. Inform. Theory*, vol. 41, no. 5, pp. 613–627, May 1995.

[10] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Stat. Assoc.*, vol. 90, pp. 1200–1224, Dec. 1995.

[11] R. R. Coifman and D. L. Donoho, "Translation-invariant de-noising," in *Wavelet and Statistics*, A. Antoniadis and G. Oppenheim, Eds. Berlin, Germany: Springer-Verlag, 1995.

[12] Y. Xu, J. B. Weaver, D. M. Healy, and J. Lu, "Wavelet transform domain filters: a spatially selective noise filtration technique," *IEEE Trans. Image Process.*, vol. 3, no. 11, pp. 747–758, Nov. 1994.

[13] Q. Pan, L. Zhang, G. Dai, and H. Zhang, "Two denoising methods by wavelet transform," *IEEE Trans. Signal Process.*, vol. 47, no. 12, pp. 3401–3406, Dec. 1999.

[14] P. Bao and L. Zhang, "Noise reduction for magnetic resonance images via adaptive multiscale products thresholding," *IEEE Trans. Medical Imaging*, vol. 22, no. 9, pp. 1089–1099, Sep. 2003.

[15] L. Zhang and P. Bao, "Edge detection by scale multiplication in wavelet domain," *Pattern Recognit. Lett.*, vol. 23, pp. 1771–1784, 2002.

[16] A. Chambotle, R. A. Devore, N. Y. Lee, and B. J. Lucier, "Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. Image Process.*, vol. 7, no. 7, pp. 319–335, Jul. 1998.

[17] M. K. Mihçak, I. Kozintsev, K. Ramchandran, and P. Moulin, "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Signal Process. Lett.*, vol. 6, no. 12, pp. 300–303, Dec. 1999.

[18] X. Li and M. Orchard, "Spatially adaptive image denoising under overcomplete expansion," in *Int. Conf. Image Process.*, Vancouver, Canada, Sep. 2000, pp. 300–303.

[19] B. M. Sadler and A. Swami, "Analysis of multiscale products for step detection and estimation," *IEEE Trans. Inform. Theory*, vol. 45, no. 4, pp. 1043–1051, April 1999.

[20] J. M. Shapiro, "Embedded image coding using zerostrees of wavelet coefficients," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3445–3462, Dec. 1993.

[21] X. Wu, "Lossless compression of continuous-tone images via context selection, quantization, and modeling," *IEEE Trans. Image Process.*, vol. 6, no. 5, pp. 656–664, May 1997.
[22] Y. Yoo, A. Ortega, and B. Yu, "Image subband coding using context-based classification and adaptive quantization," *IEEE Trans. Image Process.*, vol. 8, no. 12, pp. 1702–1715, Dec. 1999.
[23] S. T. Hsiang, "Embedded image coding using zeroblocks of subband/wavelet coefficients and context modeling," in *Proc. Data Compression Conf.*, 2001, pp. 83–92.
[24] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 42, no. 4, pp. 886–902, Apr. 1998.
[25] G. Fan and X. G. Xia, "Improved hidden Markov models in the wavelet-domain," *IEEE Trans. Signal Process.*, vol. 49, no. 1, pp. 115–120, Jan. 2001.
[26] ——, "Image denoising using local contextual hidden Markov model in the wavelet domain," *IEEE Signal Process. Lett.*, vol. 8, no. 5, pp. 125–128, May 2001.
[27] J. Liu and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Trans. Image Process.*, vol. 10, no. 11, pp. 1647–1658, Nov. 2001.
[28] ——, "Image denoising based on scale-space mixture modeling for wavelet coefficients," in *Proc. ICIP'99*, Kobe, Japan, Oct. 1999, pp. I.386–I.390.
[29] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Adaptive wiener denoising using a gaussian scale mixture model in the wavelet domain," in *Proc. 8th Int. Conf. Image Processing*, vol. 2, Thessaloniki, Greece, Oct. 2001, pp. 37–40.
[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[31] E. W. Karmen and J. K. Su, *Introduction to Optimal Estimation*. London, U.K.: Springer-Verlag, 1999.

Lei Zhang was born in 1974 in China. He received the B.S. degree in 1995 from Shenyang Institute of Aeronautic Engineering, Shenyang, China, the M.S. and Ph.D. degrees in electrical engineering from Northwestern Polytechnical University, Xi'an, China, respectively, in 1998 and 2001.

From 2001 to 2002, he was a Research Associate in the Deptartment of Computing, the Hong Kong Polytechnic University, Hong Kong. Currently, he works as a Postdoctoral in the Department of Electrical and Computer Engineering, McMaster University, Ontario, Canada. His current research interest is optimal estimation theory.

Paul Bao received the Ph.D. degree from the University of Calgary, Calgary, AL, Canada, in 1988.

He served on the faculty of the Computer Science Department, University of Calgary from 1988 to 1990 and then proceeded to work at IBM Canada as a Staff Analyst. He was an Associate Professor in the Computing Department, The Hong Kong Polytechnic University, Hong Kong, and the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong (1995–2003). He is currently a Professor in the School of Engineering, Nanyang Technological University. His research interests are in computer graphics and image processing, image-based rendering, distributed graphics and rendering, image/video coding, etc. He has published over 100 research papers in those areas.

Xiaolin Wu received the B.Sc. degree from Wuhan University, Wuhan, China, in 1982, and the Ph.D. degree from the University of Calgary, Calgary, AL, Canada, in 1988, both in computer science.

He is currently a Professor at the Department of Electrical and Computer Engineering, McMaster University, ON, Canada, and a Research Professor of Computer Science, Polytechnic University, Brooklyn, NY, and holds the National Sciences and Engineering Council of Canada Research Chair in Digital Cinema. His research interests include image processing, multimedia coding and communications, data compression, and signal quantization. He has published over one hundred research papers in these fields.