Spatially adaptive wavelet thresholding with context modeling for image denoising

S. Grace Chang, Bin Yu, and Martin Vetterli

This work was supported in part by the NSF Graduate Fellowship and U.C. Dissertation Fellowship (Chang), and in part by ARO Grant DAAH04-94-G-0232 and NSF Grant DMS-9322817 (Yu). S.G. Chang is with the Department of Electrical Engineering and Computer Science, and B. Yu is with the Department of Statistics, both at the University of California, Berkeley. M. Vetterli is with the Laboratory of Audiovisual Communications, Swiss Federal Institute of Technology (EPFL), in Lausanne, Switzerland, and is affiliated with the Department of Electrical Engineering and Computer Science at U.C. Berkeley.

Part of this work will be presented at the *IEEE International Conference on Image Processing (ICIP)* in Chicago, IL, USA, October 1998.

Abstract

The method of wavelet thresholding for removing noise, or denoising, has been researched extensively due to its effectiveness and simplicity. Much of the literature has focused on developing the best uniform threshold or best basis selection. However, not much has been done to make the threshold values adaptive to the spatially changing statistics of images. Such adaptivity can improve the wavelet thresholding performance because it allows additional local information of the image (such as the identification of smooth or edge regions) to be incorporated in the algorithm. This work proposes a spatially adaptive wavelet thresholding method based on context modeling, a common technique used in image compression to adapt the coder to changing image characterstics. Each wavelet coefficient is modeled as a random variable with the Generalized Gaussian distribution with an unknown parameter. Context modeling is used to estimate the parameter for each coefficient, which is then used to adapt the thresholding strategy. This spatially adaptive thresholding is extended to the overcomplete wavelet expansion, which yields better results than the orthogonal transform. Experimental results show that spatially adaptive wavelet thresholding yields significantly superior image quality and lower MSE than the best uniform thresholding with the original image assumed known.

I. Introduction

In this paper we address the classical problem of removing additive noise from a corrupted image, or *denoising*. In recent years there has been a plethora of work on using *wavelet thresholding* [5] for denoising in both the signal processing and statistics community, due to its effectiveness and simplicity. In its most basic form, this technique denoises in the orthogonal wavelet domain, where each coefficient is *thresholded* by comparing against a threshold; if the coefficient is smaller than the threshold, it is set to zero, otherwise it is kept or modified. The intuition is that because the wavelet transform is good at energy compaction, small coefficients are more likely due to noise, and large coefficients due to important signal features (such as edges). The threshold thus acts as an oracle deciding whether or not to keep the coefficients. Most of the literature thus far has concentrated on developing threshold selection methods, with the threshold being uniform or at best one threshold for each subband. Very little has been done on developing thresholds that are adaptive to different spatial characteristics. Other works investigate the choice of wavelet basis or expansion for the thresholding framework. One particularly interesting result is that thresholding in a shift-invariant expansion (dubbed *translation-invariant (TI) denoising* by Coifman and Donoho [3]) eliminates some of the unpleasant artifacts introduced by modifying the coefficients of the orthogonal wavelet expansion. In this paper, we use the wisdom that thresholding in a shift-invariant, overcomplete representation outperforms the orthogonal basis, and also investigate an issue that has not been explored, namely, the spatial adaptivity of the threshold value.

To motivate spatially adaptive thresholding, consider the example in Figure 1, where a square function has been corrupted by additive noise, and the goal is to recover the original function. The wavelet coefficients of the original and the noisy function are displayed in Figure 1(a). The noisy coefficients are (soft) thresholded by a single threshold in Figure 1(b), and one can see that, especially in the finest scale, there are some coefficients corresponding to noise which have not been set to zero, and that some of these noisy coefficients are larger in magnitude than those coefficients corresponding to the signal. Thus, with a uniform threshold, it may not be feasible to have both the benefits of keeping the important signal features and killing the noisy coefficients. On the other hand, one can reap both benefits with adaptive thresholds by choosing the threshold value to be very small in the regions of the peaks due to the step function, and large otherwise (Figure 1(c))[1] Thus the question becomes, how one distinguishes between the coefficients that are mainly due to signal and those mainly due to noise. Also, how should the thresholds be adjusted? These are the questions that we will answer in this paper with our proposed algorithm.

Most natural images have non-stationary properties, since they typically consist of regions of smoothness and sharp transitions. These regions of varying characteristics can be well differentiated in the wavelet domain, as can be seen in the wavelet decomposition of the *lena* image in Figure 2. One observes areas of high and low energy (or large and small coefficient magnitude), represented by white and black pixels, respectively. Areas of high energy correspond to signal features of sharp variation such as edges and textures; areas of low energy correspond to smooth regions. When noise is added, it tends to increase the magnitude of the wavelet coefficient on average. Specifically, in smooth regions, one expects the coefficients to be dominated by noise, thus most of these coefficients should be removed, especially since noise is highly visible here. In regions of sharp transition, the coefficients have a lot of energy due to the signal, and some due to noise (which is not as visible in these regions), thus they should be kept, or at most modified only a little, to ensure that most of the signal details are retained. Thus, the idea is to distinguish

[1]For the sake of illustrating the effectiveness of varying thresholds, the regions of the true peaks are assumed to be known in this example.

between the low and high energy regions, and modify the coefficients using a spatially adaptive thresholding strategy.

To accomplish spatial adaptive thresholding, we model each wavelet coefficient as a realization from a given probability distribution, whose parameter is to be estimated. This parameter in turn is used to find the appropriate threshold. It has long been accepted in the subband coding community that for a large class of images, the coefficients in each subband form a distribution well described by the Generalized Gaussian distribution (GGD) [11]. Instead of using one parameter for each subband level, several wavelet-based image coders have achieved better performances by modeling the wavelet coefficients as a mixture of GGD with unknown slowly spatially varying parameters [7], [12]. The estimation of the parameter for a given coefficient is conditioned on a function of its neighboring coefficients, a method called *context-modeling* frequently used in compression for differentiating pixels of varied characteristics and adapting the coder. Context modeling allows one to group pixels of similar nature but not necessarily spatially adjacent, and to gather statistical information from these pixels. Now, given that one can estimate the parameter for each coefficient, the next step is to use them to calculate the threshold. The work in [2] found that when the signal coefficients are modeled as Generalized Gaussian random variables and the noise as Gaussian, the threshold $\tilde{T} = \sigma_n^2/\sigma_x$ is a good approximation to the optimal threshold which minimizes the mean squared error of the soft-thresholding estimator, where σ_n^2 is the noise power, and σ_x is the standard deviation of the signal. The simplicity of this threshold makes it easy to achieve spatial adaptivity – one uses context modeling to quantify the local characteristic in σ_x, which in turn yields a threshold $\tilde{T}$ adaptive on a pixel-by-pixel manner.

Our proposed algorithm is based on using adaptive thresholding in the overcomplete wavelet expansion. It outperforms both using only adaptive thresholding in the orthogonal expansion or using only uniform thresholding in the overcomplete expansion like the TI denoising. That is, by combining spatially adaptive thresholding and overcomplete expansion, we achieve results which are significantly more superior than either method alone. Firstly, the adaptive threshold selection does a good job at removing noise in smooth regions, while not disturbing too much the edge and texture regions. Thresholding in the overcomplete expansion acts as an additional averaging which further attenuates the remaining noise.

The organization of this paper is as follows. In Section II, we introduce the threshold selection method when there is only one class of Generalized Gaussian distributed random variables corrupted by independent additive Gaussian noise. Because this threshold selection is based on *iid* noise assumption, the discussion will first set in the orthogonal wavelet transform. Then context modeling is introduced to allow coefficients be modeled as random variables of different parameters, and the parameters are used to make the threshold spatially adaptive. Finally, we discuss how to extend this adaptive method in the orthogonal expansion to the overcomplete expansion. In Section III, we will compare the spatially adaptive results with those from the best uniform thresholding strategy (in the mean squared error sense, and based on knowing the original image), in both the orthogonal and overcomplete expansion. Results will show that the combination of using spatially adaptive thresholding and overcomplete expansion yields significantly superior results in both the visual quality and the mean squared error.

II. Adaptive Algorithm

The adaptive algorithm will be developed in the following manner. Firstly, the concept of wavelet thresholding in the orthogonal wavelet transform is introduced, and coefficients in each subband are modeled as realizations of one class of GGD (with unknown parameter), corrupted by additive white Gaussian noise. We then describe the threshold selection method developed in [2] which achieves near-optimal soft-thresholding under these distributions. To make this thresholding approach spatially adaptive, *each coefficient* (rather than each subband) is modeled as a GGD random variable with a different unknown parameter which is estimated on a pixel level via context modeling. This spatial mixture of distributions allows the image characteristics to be quantified locally in the distribution parameters, which are then used to adjust the threshold for each coefficient. Lastly, since the aforementioned algorithm is developed in the orthogonal expansion where the coefficients are uncorrelated, the algorithm will need to be modified to extend to the overcomplete expansion where coefficients are correlated.

A. Coefficient Modeling and Threshold Selection

Let the corrupted $N \times N$ image be denoted by

$$y[i,j] = f[i,j] + \varepsilon[i,j], i,j = 1,\ldots,N,$$

where $\{f[i,j]\}$ is the original image, and $\{\varepsilon[i,j]\}$ are *iid* $N(0,\sigma_n^2)$ and independent of $\{f[i,j]\}$. The goal is to recover $\{f[i,j]\}$ from the noisy observations $\{y[i,j]\}$. The image size N is assumed to be power of 2 for simplicity.

To accomplish wavelet thresholding for denoising, the observations $\{y[i,j]\}$ are first transformed into the wavelet domain. The necessary notations for the wavelet transform will be introduced here, and the readers are referred to references such as [8], [10] for more details. The 2D discrete orthogonal wavelet transform (DWT) can be implemented as a critically sampled octave-band filter bank, where separable filtering is used. It is often convenient to cluster these coefficients into groups or *subbands* of different scales and orientations as in Figure 4, where, for example, the label HL_1 refers to those coefficients at the first scale of decomposition which are the output of the highpass filter in the horizontal direction and the lowpass filter in the vertical direction. Let $Y^{(s,o)}[i,j], i,j = 1, \dots, N/2^s$, denote the wavelet coefficients of $\{y[i,j]\}$ at a particular scale s and orientation o, where $s = 1, 2, \dots, J$ and $o \in \{HL, LH, HH, LL\}$.

It has been observed that for a large class of images, the coefficients from each subband (except LL) form a symmetric distribution that is sharply peaked at zero [11], well described by the zero-mean Generalized Gaussian distribution (GGD)

$$GG_{\beta,\sigma_x}(x) = C(\beta,\sigma_x)\, e^{-(\alpha(\beta,\sigma_x)|x|)^\beta} \tag{1}$$

where

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

and $\Gamma(t) = \int_0^\infty e^{-u} u^{t-1} du$ is the gamma function. The parameter σ_x is the standard deviation and determines the spread of the density function, and the parameter β is called the *shape parameter*. Special cases of the GGD density function include the Gaussian distribution with $\beta = 2$, and the Laplacian with $\beta = 1$. Let $\{X^{(s,o)}[i,j]\}$ and $\{V^{(s,o)}[i,j]\}$ denote the wavelet coefficients of the original signal $\{f[i,j]\}$ and the noise $\{\varepsilon[i,j]\}$, respectively. For each subband, the signal coefficients $\{X^{(s,o)}[i,j]\}$ are modeled as independent samples of distribution $p_X(x) = GG_{\beta,\sigma_x}(x)$, and since the wavelet transform is orthogonal, the noise coefficients are independent samples of the Gaussian distribution $p_V(v) = \phi(v,\sigma_n^2) = 1/\sqrt{2\pi\sigma_n^2}\exp\{-\frac{v^2}{2\sigma_n^2}\}$. Let the estimator be restricted to the *soft-threshold* estimator of the form $\hat{X}^{(s,o)}[i,j] = \eta_T(Y^{(s,o)}[i,j])$, where $\eta_T(x) = \mathrm{sgn}(x)\cdot\max(|x|-T,0)$ is the soft-thresholding function, and the optimal threshold T^* is defined to be the argument which minimizes the expected squared error,

$$T^* = \arg\min_T E_{Y|X,X}(\eta_T(Y) - X)^2 \tag{2}$$

where $Y|X \sim \phi(y-x,\sigma_n^2)$ and $X \sim GG_{\beta,\sigma_x}(x)$. In [2], it was found that at least for the range of $\beta \in [0.8, 4]$, T^* can be well approximated by $\tilde{T} = \sigma_n^2/\sigma_x$, with at most 5% difference from the minimum expected squared error. Because the threshold $\tilde{T}$ depends only on the standard deviation σ_x and not on the shape parameter β, it may not yield a good approximation for other values of β than the range tested here, and the threshold may need to be modified to incorporate β. However, since in practice the values of $\beta = 1, 2$ are typically used in modeling wavelet coefficients of real images, well within the range of β tested here, the simple form of the threshold $\tilde{T}$ is appropriate for our purpose. The curve of expected squared error is very flat near the optimal threshold T^*, implying that the error is not very sensitive to a slight perturbation near T^*.

The threshold $\tilde{T} = \sigma_n^2/\sigma_x$ is not only nearly optimal but also has an intuitive appeal. For such a choice, the normalized threshold $\tilde{T}/\sigma_n$ is inversely proportional to σ_x, the standard deviation of X, and proportional to σ_n, the noise standard deviation. When $\sigma_n/\sigma_x \ll 1$, the signal is much stronger than the noise, thus $\tilde{T}/\sigma_n$ is chosen to be small in order to preserve most of the signal and remove some of the noise; vice versa, when $\sigma_n/\sigma_x \gg 1$, the noise dominates and the normalized threshold is chosen to be large to remove the noise which has overwhelmed the signal.

The proposed threshold $\tilde{T}$ can easily be adjusted to the signal and noise energy as reflected in σ_x and σ_n. By estimating the parameter σ_x for each subband, we have a uniform threshold $\tilde{T}$ adaptive on a subband-level. However, better denoising performance can be achieved by using spatially adaptive thresholds, whose derivation will be described in the following section.

B. Context Modeling for Spatial Adaptivity

To achieve a spatially adaptive thresholding strategy, the wavelet coefficients are modeled as components in a discrete random field, with a collection of independent zero-mean GGD random variables whose parameters β and σ_x are spatially varying. The parameter σ_x needs to be estimated for each coefficient to make the threshold $\tilde{T}$ spatially adaptive. This can be accomplished by *context modeling*, an idea used frequently in image compression for adapting the coder to changing image characteristics. That is, the statistical model for a given coefficient is conditioned on a function of its neighbors. Several model-based coders have utilized information from causal quantized neighbors to determine the context model for each coefficient [7], [12]. The coder in [7] estimates the shape and standard variation parameters by the maximum likelihood (ML) estimator from the quantized coefficients within a causal neighborhood, allowing essentially an infinite mixture of distributions[2]. In the wavelet-based compression scheme in [12], context modeling was used to categorize coefficients into several classes of Laplacian distributions with different values of σ_x. The conditioning was based on the weighted average of the coefficient magnitude in a causal neighborhood, and each class was formed by clustering coefficients whose associated weighted averages fall within a specified range. The distribution parameter is estimated from the coefficients for each class, which is then used to adapt the coder. Since the parameter and the description of each class need to be sent as overhead, only four classes were used in [12]. For the denoising problem, there is no need to conserve bits, thus it is not necessary to explicitly classify the pixels, and parameters can be estimated for each coefficient via a moving window, resulting in virtually an infinite mixture of distributions.

Consider one particular subband with M^2 coefficients, $\{Y^{(s,o)}[i,j]\}$. To simplify notation, we drop the superscript (s,o), and use it only when necessary for clarity. Each coefficient $Y[i,j]$ is modeled as a random variable whose variance can be estimated as follows. Consider a neighborhood of $Y[i,j]$, and the absolute value of its p elements are placed in a $p \times 1$ vector $\boldsymbol{u}_{ij}$. One possible choice is the eight nearest neighbors of $Y[i,j]$ in the same subband, plus its parent coefficient $Y^{(s+1,o)}[\lceil i/2 \rceil, \lceil j/2 \rceil]$ (see Figure 4 for the definition of parent-child relationship). To characterize the activity level of the current pixel, we calculate a weighted average of the absolute value of the neighbors,

$$Z[i,j] = \boldsymbol{w}^t \; \boldsymbol{u}_{ij}.$$

The weights $\boldsymbol{w}$ is found by using the least squares estimate, that is,

$$\begin{aligned} \boldsymbol{w}^* &= \arg\min_{\boldsymbol{w}} \sum_{i,j} (|Y[i,j]| - \boldsymbol{w}^t \; \boldsymbol{u}_{ij})^2 \\ &= (\boldsymbol{U}^t \boldsymbol{U})^{-1} \boldsymbol{U}^t \, |\boldsymbol{Y}| \end{aligned}$$

where $\boldsymbol{U}$ is a $M^2 \times p$ matrix with each row being $\boldsymbol{u}_{ij}^t$, for all i,j, and $\boldsymbol{Y}$ is the $M^2 \times 1$ vector containing all coefficients $Y[i,j]$. Notice that the absolute values of the neighbors rather than their original values are used in the averaging. This is because orthogonal wavelet coefficients are uncorrelated, and thus an average of the neighbors does not yield much information about the coefficient of interest. However, the absolute value or the squared values of neighboring

[2] To be exact, the shape parameter are restricted to be a member of a discrete set. The quantizer is pre-designed for a fixed discrete set of shape parameter β and slope λ/σ_x^2 where λ is determined by the target rate, so that the coding can be done quickly with the aid of a table lookup.

coefficients are correlated, and therefore their averages are useful in collecting information about other coefficients in the near vicinity.

The variance of coefficient $Y[i,j]$ is estimated from other coefficients whose context lie in the interval $[Z[i,j]-\Delta_1, Z[i,j]+\Delta_2]$. To develop an intuition for this, it is helpful to examine Figure 5, which plots the pairs $\{Z[i,j], Y[i,j]\}, i,j = 1,\ldots,M$. The points are clustered within a cone shape centered at origin. Taking an interval of small valued $Z[i,j]$, the associated coefficients $\{Y[i,j]\}$ have a small spread; on the other hand, an interval of large valued $Z[i,j]$ has corresponding $\{Y[i,j]\}$ with a larger spread (the intervals are of different widths to capture the same number of points). This suggests that the context provides a good indication of local variability. Thus, for a given coefficient $Y[i_0,j_0]$, we place an interval around $Z[i_0,j_0]$, and the variance of $Y[i_0,j_0]$ is estimated from the points $Y[i,j]$ whose context falls within this window. In particular, we take L closest points above $Z[i_0,j_0]$ and L closest points below, resulting in a total of $2L+1$ points, where we choose $L = \max(50, M^2/10)$ to ensure that enough points are used to estimate the variance. Note that this is a moving window rather than the fixed classes in [12], and thus allows a continuous range of estimate values. Let $\mathcal{B}_{i_0j_0}$ denote the set of points $\{Y[i,j]\}$ whose context falls in the moving window. The estimate of the variance $\sigma_x^2[i_0,j_0]$ is then

$$\hat{\sigma}_x^2[i_0,j_0] = \max\left(\frac{1}{2L+1}\sum_{[k,\ell]\in\mathcal{B}_{i_0j_0}} Y[k,\ell]^2 - \sigma_n^2, 0\right). \quad (3)$$

The term σ_n^2 needs to be subtracted because $\{Y[i,j]\}$ are the noisy observations, and the noise is independent of the signal, with variance σ_n^2. The threshold at location $[i_0,j_0]$ is then

$$\tilde{T}_{i_0j_0} = \frac{\sigma_n^2}{\hat{\sigma}_x[i_0,j_0]}.$$

Calculating the threshold $\tilde{T}_{ij}$ for every location $[i,j]$ yields a spatially adaptive threshold. In the implementation, the context $\{Z[i,j]\}$ are first sorted, and a moving window is placed over them, so the set $\mathcal{B}_{ij}$ and the variance estimate $\hat{\sigma}_x^2[i,j]$ can be updated efficiently.

Up to now we have not discussed how to estimate the noise variance σ_n^2. In some practical cases, it is possible to measure σ_n^2 based on information other than the corrupted observation. If this is not the case, we estimate it by using the robust median estimator in the highest subband of the wavelet transform,

$$\hat{\sigma}_n = \text{Median}(|Y[i,j]|)/.6745\,, \quad Y[i,j] \in \text{ subband } HH_1,$$

also used in [5].

C. Alternative Methods

There are several other possible alternative approaches which will be discussed below.

1. One may ask why the local variance is not estimated from, say, a local window (as in [7]), but rather from an indirect way of grouping the coefficients first via its context. Estimating from a local neighborhood is simple, and, as demonstrated by the good performance of the image coder in [7], it yields an estimate good enough for adapting the coder. However, our experience with noisy images show that such an estimate yields considerably more unreliable variance estimates, σ_x^2, and also blotchy denoised image. This is because the estimate is highly sensitive to the window size we choose: a small window contains few points and thus yields unreliable estimates; a large window adapts slowly to different characteristics. The context-based grouping allows one to congregate those coefficients with similar context though not necessarily spatially adjacent. It also allows a large number of coefficients to be used in the variance estimation, thus yielding a more reliable estimate. Simulations show that performance is not sensitive to the neighborhood choice $\mathcal{B}_{ij}$ and the weight $\boldsymbol{w}$ used in the context calculation, as a simple equally weighted average of the eight nearest neighbors in the same subband yield approximately the same result.

2. The method we have proposed is a *two-pass* process: the first pass calculates the weighted average $\{Z[i,j]\}$ of the absolute values of the neighboring *noisy* coefficients, and then $\{Z[i,j]\}$ are sorted; the second pass collects the noisy coefficients with similar values of $Z[i,j]$, estimates the signal variance, σ_x^2, from the noisy coefficients and then the thresholds for the thresholding function. It is worthwhile to investigate the algorithm performance when the context modeling and the parameter estimation are performed on the *denoised* coefficients instead, since, intuitively, if the coefficients are really denoised, they should yield more reliable information. This simple intuition is, however, not as straightforward to implement as it seems. To do this in a two-pass algorithm is difficult, since $Z[i,j]$ is a weighted average of neighboring denoised coefficients, but the threshold used to denoise these coefficients are estimated from other denoised coefficients with similar context. A simple-minded alternative solution is to use a *one-pass* modification of our algorithm, where the conditioning and estimation are based on *causal, denoised* coefficients, much along the same philosophy as one-pass compression methods conditioning on causal quantized data [7], [12]. Assume a scanning order of row by row, and initialize the first coefficient as already denoised, that is, $\hat{X}[1,1] = Y[1,1]$. For every new coefficient at location $[i,j]$, the context is conditioned on $Z[i,j] = \boldsymbol{w}^t \boldsymbol{u}_{ij}$ where $\boldsymbol{u}_{ij}$ is now the vector containing the absolute

value of denoised coefficients $\hat{X}[i,j]$ in a causal neighborhood, and the elements of $\boldsymbol{w}$ are simply the equal weights (these choices are made for simplicity since the denoising performance is not too sensitive to the neighborhood selection and weight vector $\boldsymbol{w}$.) The GGD parameter $\sigma_x[i,j]$ is estimated from past denoised coefficients whose contexts are similar, and $2L+1$ coefficients are used (or all of the available coefficients so far if less than $2L+1$ coefficients have been denoised.) Since the coefficients are already denoised, the estimation of $\sigma_x[i,j]$ is $\hat{\sigma}_x^2[i,j] = \frac{1}{2L+1}\sum_{[k,\ell]\in\mathcal{B}_{ij}} \hat{X}[k,\ell]^2$ instead of (3). Simulations show this approach to run into problems especially when the noise power σ_n^2 is large, causing many coefficients to be denoised to zero. Having too many consecutive zero coefficients is likely to cause $\hat{\sigma}_x[i,j]$ to be zero, which then translates to an infinite threshold (i.e. $Y[i,j]$ is thresholded to zero). This in turn may cause all the subsequent coefficients to be thresholded to zero. This phenomenon is frequently encountered in backward adaptive compression methods which adapts based on causal quantized coefficients: a run of zero coefficients may cause all subsequent coefficients to be quantized to zero as well. Some work ameliorate this problem by looking ahead to identify *unpredictable sets*, coefficients whose neighbors are zero, but who should not be quantized to zero [7], [12]. This logic can be applied to the denoising framework as well. When the algorithm computes $Z[i,j]$, it identifies the locations $[k,\ell]$ in the causal neighborhood $\mathcal{B}_{ij}$ for which $\hat{X}[k,\ell]=0$ but $|Y[k,\ell]|\geq\sigma_n$, and these $Y[k,\ell]$ are substituted for the zero $\hat{X}[k,\ell]$ to be used in the computation of $Z[i,j]$ and $\hat{\sigma}_x[i,j]$. Simulations show the resulting images to yield slightly worse MSEs than the previously proposed method, but they are visually considerably more noisy because the variance estimate $\hat{\sigma}_x[i,j]$ is noisier to begin with.

Another variation is to use the denoised coefficient for context modeling, but the observed noisy coefficients for estimating $\sigma_x^2[i,j]$ as in (3). Again, without taking some caution about the runs of zero coefficients, the variance estimate may be inadequate for several rows (recall the scanning is row by row) before having enough non-zero causal neighbors for collecting valid information. The denoised images are also similar to the ones described above, having slightly worse MSEs than the proposed two-passed algorithm, but are visually more noisy.

3. A central part of our spatially adaptive algorithm is based on modeling the variance $\sigma_x^2[i,j]$ to be non-constant and varying throughout the image. This is reminiscent of the *heteroscedasticity*, or non-constant variance, problem in statistics. Let $\{Y[i,j]\}$ be the observed noisy wavelet coefficients, and each $Y[i,j]$ a random variable whose variance $\gamma^2[i,j]$ is non-constant. A common

approach to the heteroscedasticity problem is to model $\gamma^2[i,j]$ as a function of some design vector, $\boldsymbol{u}_{ij}$. Traditionally there are two approaches in estimating this function, parametric and non-parametric. Since we have an assumed distribution on the wavelet coefficients (i.e. GGD), the parametric approach will be used here. The readers are referred to [1], [9] and related literatures for more details on heteroscedasticity models. Using a parametric function to describe the variance $\gamma^2[i,j]$ has the advantage that it allows a compact representation of the non-constant variance, useful for image analysis and understanding. In contrast, although the non-parametric approach described in II-B works well, it does not lend itself to any tractable analysis.

In the previous section, we have described the noisy coefficient $Y[i,j]$ as a sum of two random variables, $X[i,j] \sim$ GGD and $V[i,j] \sim$ Gaussian. Unless $X[i,j]$ is a Gaussian distributed random variable, there is no closed form expression for the distribution of $Y[i,j]$. However, often one observes the wavelet coefficients for images to be sharply peaked at zero, better described by the Laplacian density function. Furthermore, the noisy coefficients also form a histogram which is sharply peaked at zero. Thus, for simplicity and for the sake of tractable analysis, we assume the noisy coefficient $Y[i,j]$ to be Laplacian distributed, or, alternatively, $|Y[i,j]|$ be exponentially distributed. Recall that in our context modeling framework, the design vector $\boldsymbol{u}_{ij}$ at location $[i,j]$ is the vector containing the absolute value of the neighboring (noisy) coefficients (see Section II-B), and the variance for $Y[i,j]$ is estimated by collecting other coefficients with similar values of $\boldsymbol{w}^t\boldsymbol{u}_{ij}$. Formally, our heteroscedasticity model is

$$|Y[i,j]| \sim \frac{1}{\gamma[i,j]} e^{y/\gamma[i,j]}, y \geq 0,$$

where the standard deviation is

$$\gamma[i,j] = K_{\boldsymbol{\theta}}(\boldsymbol{w}^t\boldsymbol{u}_{ij}),$$

$K_{\boldsymbol{\theta}}(\cdot)$ is a smooth function such as a polynomial of order r, with unknown parameter $(r+1)\times 1$ vector $\boldsymbol{\theta}$, and $\boldsymbol{w}$ is the regression parameter (i.e. the weights for the weighted average of the neighboring coefficients contained in $\boldsymbol{u}_{ij}$). Modeling $\gamma[i,j]$ as a function of $\boldsymbol{w}^t\boldsymbol{u}_{ij}$ can be justified by observing that the plot of $\{(\boldsymbol{w}^t\boldsymbol{u}_{ij}, Y[i,j]\}_{i,j}$ often resides within a cone shape (see Figure 5), implying that the variability of $Y[i,j]$ highly depends on $\boldsymbol{w}^t\boldsymbol{u}_{ij}$.

To estimate the parameters $\boldsymbol{\theta}$ and $\boldsymbol{w}$, we use the likelihood approach. The negative log-likelihood of $|Y[i,j]|$ is

$$\log K_{\boldsymbol{\theta}}(\boldsymbol{w}^t\boldsymbol{u}_{ij}) + \frac{|Y[i,j]|}{K_{\boldsymbol{\theta}}(\boldsymbol{w}^t\boldsymbol{u}_{ij})}.$$

For $\{|Y[i,j]|\}_{i,j=1,\ldots,N}$, the negative log-likelihood, or the *likelihood function*, is

$$L(\boldsymbol{\theta}, \boldsymbol{w}) = \sum_{i,j=1}^{N} \left(\log K_{\boldsymbol{\theta}}(\boldsymbol{w}^t \boldsymbol{u}_{ij}) + \frac{|Y[i,j]|}{K_{\boldsymbol{\theta}}(\boldsymbol{w}^t \boldsymbol{u}_{ij})} \right).$$

The likelihood function is minimized over both parameters $\boldsymbol{\theta}$ and $\boldsymbol{w}$ to find their optimal values. One way to do this is to start with an initial $\boldsymbol{w}$ being the linear least squares estimate, $\hat{\boldsymbol{w}}_{LS} = \arg\min_{\boldsymbol{w}} \sum_{i,j} (|Y[i,j]| - \boldsymbol{w}^t \boldsymbol{u}_{ij})^2$. Then $\boldsymbol{\theta}$ is estimated as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \hat{\boldsymbol{w}}_{LS}).$$

The regression parameter $\boldsymbol{w}$ is refined one step further as

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} L(\hat{\boldsymbol{\theta}}, \boldsymbol{w}).$$

After obtaining $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{\theta}}$, the standard deviation of $Y[i,j]$ is estimated by $\hat{\gamma}[i,j] = K_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{w}}^t \boldsymbol{u}_{ij})$, and the variance estimate of the clean coefficient $X[i,j]$ is $\hat{\sigma}_x^2[i,j] = \max(0, \hat{\gamma}^2[i,j] - \hat{\sigma}_n^2)$. The threshold is then calculated as before to be $\tilde{T}_{ij} = \hat{\sigma}_n^2 / \hat{\sigma}_x[i,j]$.

Polynomials of order $r = 1, 2$ were experimented with. Simulations show this parametric estimation of $\gamma^2[i,j]$ to differentiate well between regions of high energy (e.g. edges and textures) and smooth areas. That is, the variance estimate is larger in the edge and texture region, and smaller in the smooth regions. However, these values are not appropriate since the subsequently calculated variance estimate of $X[i,j]$, $\hat{\sigma}^2[i,j]$, results in zero in many subbands, which in turn translates to killing all the coefficients in the thresholding. This phenomenon may be due to the disparity between the parametric modeling of the non-constant variance and the noisy observation modeling: in the parametric approach, the observed noisy coefficients are Laplacian distributed, whereas in the original model, the observations are sums of a Laplacian and Gaussian random variable. Nevertheless, the likelihood approach to the heteroscedasticity problem may be valuable to other analysis and applications.

D. Thresholding in Overcomplete Expansion

Thresholding in the orthogonal wavelet domain has been observed to produce significantly noticeable artifacts such as Gibbs-like ringing and blips. To ameliorate this unpleasant phenomenon, Coifman and Donoho [3] proposed the *translation-invariant (TI) denoising*. Let $\mathit{Shift}_{k,\ell}[y]$ denote the operation of circularly shifting the input image y by k indices in the vertical direction and ℓ indices in the horizontal, and let $Unshift_{k,\ell}[y]$ be a similar operation but in the opposite direction. Also, let $Denoise[y, T]$ denote the operation of taking the DWT of the input image y, threshold it with a chosen uniform threshold T, then transform it back to the space domain. Then TI denoising yields an output which is the average of the thresholded copies over all possible shifts: $\hat{f} = \frac{1}{N^2} \sum_{k,\ell=0}^{N-1} Unshift_{k,\ell}[Denoise[Shift_{k,\ell}[y], T]]$. The rationale is that since the orthogonal wavelet transform is a time-varying transform and thresholding the coefficients produces ringing-like phenomena, thresholding a shifted input would produce ringing at different locations, and averaging over all different shifts would yield an output with more attenuated artifacts than a single copy alone. TI denoising can be shown to be equivalent to thresholding in the overcomplete representation implemented by the *non-subsampled filter bank* as will be described below, up to some scaling in the thresholds. It has been shown to remove some of the ringing artifacts, because denoising in the redundant expansion can be interpreted as an additional averaging. Thus we proceed to extend our spatial adaptive algorithm to this redundant expansion.

The adaptive algorithm in the orthogonal basis described above can easily be extended to the overcomplete basis. Now consider the same orthogonal filters but used in a filter bank without downsampler (see Figure 6 for the 1D case, and the 2D case is achieved by simply extending the 1D filtering to separable filtering; refer to [10] for more detail on non-subsampled filter banks). The filters are renormalized by $1/\sqrt{2}$ so that coefficient energy stays the same. This decomposition is a redundant representation, and there are correlations between the decomposition coefficients. For example, at the first level of decomposition, the odd and even coefficients (in each direction) are correlated. Thus, we can separate the coefficients into four sets of uncorrelated coefficients, namely, $\{Y[2i, 2j]\}, \{Y[2i, 2j+1]\}, \{Y[2i+1, 2j]\}$ and $\{Y[2i+1, 2j+1]\}$. For the s-th level decomposition, the coefficients can be separated into 2^{2s} sets, each containing uncorrelated coefficients, and they are $\{Y[2^s i + k_1, 2^s j + k_2]\}_{i,j}, k_1, k_2 = 0, 1, \ldots, 2^s - 1$. Since each set contains uncorrelated coefficients, the noise are also *iid* within each set as well, and thus the adaptive algorithm can be used for each set of coefficients. This approach lets us still use the independent noise assumption and circumvent the issue of denoising correlated signal coefficients with correlated noise, which is not an easy task. That is, if the coefficients are correlated, then one can conceivably do better than thresholding each coefficient independently; one could look at the numerous correlated coefficients and do a *joint thresholding*. This is still an open problem and will be investigated. For correlated noise, [6] proved some minimax properties of using a modified *universal threshold*, $\sigma_n^{(s)}\sqrt{2\log N}$, where σ_s is the standard deviation of the noise at decomposition level s. The framework is for deterministic signal, however, and not the Bayesian framework used here. Thus, for simplicity, we separate the coefficients into groups of uncorrelated coefficients before using the thresholding algorithm.

There are two other minor details in this implementation. Firstly, one needs to alter the noise power σ^2 at each decomposition scale to $\sigma^2/4^s$ due to the renormalization of the filters. Secondly, the definition of the parent coefficient used in the neighborhood of the context is slightly changed: the parent of a coefficient in scale s is simply the coefficient at the same spatial location in scale $s+1$.

III. Experimental Results

We use the images *barbara* and *lena* as test images. *iid* Gaussian noise at different levels of σ^2 are generated using *randn* in MATLAB. For the orthogonal wavelet transform, four levels of decomposition are used, and the wavelet employed is Daubechies' symmlet with 8 vanishing moments [4]. There are four methods that we compare, and the MSE results are shown in Table I. The *AdaptDWT* method refers to the proposed adaptive thresholding using the orthogonal transform DWT, and *AdaptNS* refers to adaptive thresholding using the non-subsampled wavelet transform. These two are compared against the best uniform thresholding techniques (in the MSE sense) when the original uncorrupted image is assumed to be known. For thresholding with DWT, in each subband, we find the *oracle* threshold T_{orc} as

$$T_{orc} = \arg\min_T \|\eta_T(Y[i,j]) - X[i,j]\|^2$$

where $Y[i,j]$ and $X[i,j]$ are the wavelet coefficients of the noisy observation y and original image f, respectively. This method is labeled *OrcUnifDWT* in Table I. Similarly, this is extended to the non-subsampled wavelet transform, where a different threshold is found for each set of uncorrelated coefficients within each subband (thus 2^{2s} thresholds for a subband at scale s). This method is labeled *OrcUnifNS*. Figure 7 shows a magnified region in the *barbara* image for $\sigma = 25$ and the *lena* image for $\sigma = 22.5$. The *AdaptNS* method outperforms all the other methods in both visual quality and MSE performance. It yields significantly less ringing artifacts and blotchiness than the methods using DWT. The *OrcUnifNS* method using uniform thresholds in the non-subsampled framework still shows significant noise in the smooth background. Thus, it

TABLE I

COMPARING THE MSE OF THE SPATIALLY ADAPTIVE ALGORITHM WITH OPTIMAL SUBBAND UNIFORM THRESHOLD IN THE DWT AND OVERCOMPLETE EXPANSION FOR VARIOUS TEST IMAGES AND σ.

MSE/σ	12.5	15	17.5	20	22.5	25
	barbara					
AdaptDWT	61.4	78.3	94.0	111.6	127.5	144.8
OrcUnifDWT	62.2	80.7	99.2	117.3	136.8	155.0
AdaptNS	**43.5**	**56.0**	**68.7**	**83.1**	**97.5**	**112.2**
OrcUnifNS	51.2	66.3	81.0	96.7	112.0	128.2
	lena					
AdaptDWT	36.1	42.7	50.2	58.1	66.5	72.9
OrcUnifDWT	36.1	43.7	51.3	58.8	67.4	73.7
AdaptNS	**27.5**	**32.7**	**38.4**	**44.1**	**51.1**	**56.5**
OrcUnifNS	29.8	35.9	42.3	48.7	55.7	61.2

is both the spatial adaptive thresholds and the overcomplete representation that contribute to the superior quality of the *AdaptNS* method. The adaptive methods denoise better especially in the flat regions, where the uniform methods yields images with much noise and "blips". Note that although the MSEs for the *lena* image is similar between the adaptive and uniform oracle methods, the visual quality in the adaptive method is far superior as it produces a denoised image that is smooth in the flat regions and has less artifacts around the edges as well.

IV. Conclusion

We have proposed a simple and effective spatially and scale-wise adaptive method for denoising via wavelet thresholding in the overcomplete expansion. The adaptivity is based on context-modeling which enables a pixel-wise estimation of the signal variance and thus the best threshold. The issue of spatially adapting the threshold values has not been addressed in the literature. As we have shown in this paper, adapting the threshold values to local signal energy allows us to keep much of the edge and texture details, while eliminating most of the noise in smooth regions, something that may be hard to achieve with a uniform threshold. The results show substantial improvement over the optimal uniform thresholding when the original image is assumed known, both in visual quality and mean squared error.

References

[1] R.J. Carroll, "Adapting for heteroscedasticity in linear models," *Annals of Statistics*, Vol. 10, No. 4, pp1224-1233, 1982.

[2] S.G. Chang, B. Yu, and M. Vetterli, "Image Denoising via Lossy Compression and Wavelet Thresholding," *Proc. IEEE Int. Conf. Image Processing*, Vol.1, pp. 604-607, Oct. 1997.

[3] R.R. Coifman and D.L. Donoho, "Translation-invariant de-noising," Wavelets and Statistics, A. Antoniadis and G. Oppenheim eds., Springer-Verlag Lecture Notes, 1995.

[4] I. Daubechies, *Ten Lectures on Wavelets*, vol. 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, SIAM, Philadelphia, 1992.

[5] D.L. Donoho and I.M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol 81, pp. 425-455, 1994.

[6] I.M. Johnstone and B.W. Silverman, "Wavelet threshold estimators for data with correlated noise," *Journal of the Royal Statistical Society*, Ser. B, **59**, 1997.

[7] S. LoPresto, K. Ramchandran, and M. Orchard, "Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization framework," *Proc. Data Compression Conference*, Snowbird, Utah, March 1997.

[8] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Pat. Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674-693, July 1989.

[9] H.-G. Müller and U. Stadtmüller, "Estimation of heteroscedasticity in regression analysis," *Annals of Statistics*, Vol. 15, No. 2, pp. 610-625, 1987.

[10] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*, Prentice Hall, Englewood Cliffs, NJ, 1995.

[11] P.H. Westerink, J. Biemond, and D.E. Boekee, "An optimal bit allocation algorithm for sub-band coding", *Proc. Int. Conf. on Acous., Speech and Signal Process.*, Dallas, Texas, pp. 1378-1381, April 1987.

[12] Y. Yoo, A. Ortega, and B. Yu, "Image Subband Coding using Progressive Classification and Adaptive Quantization" preprint, 1997.

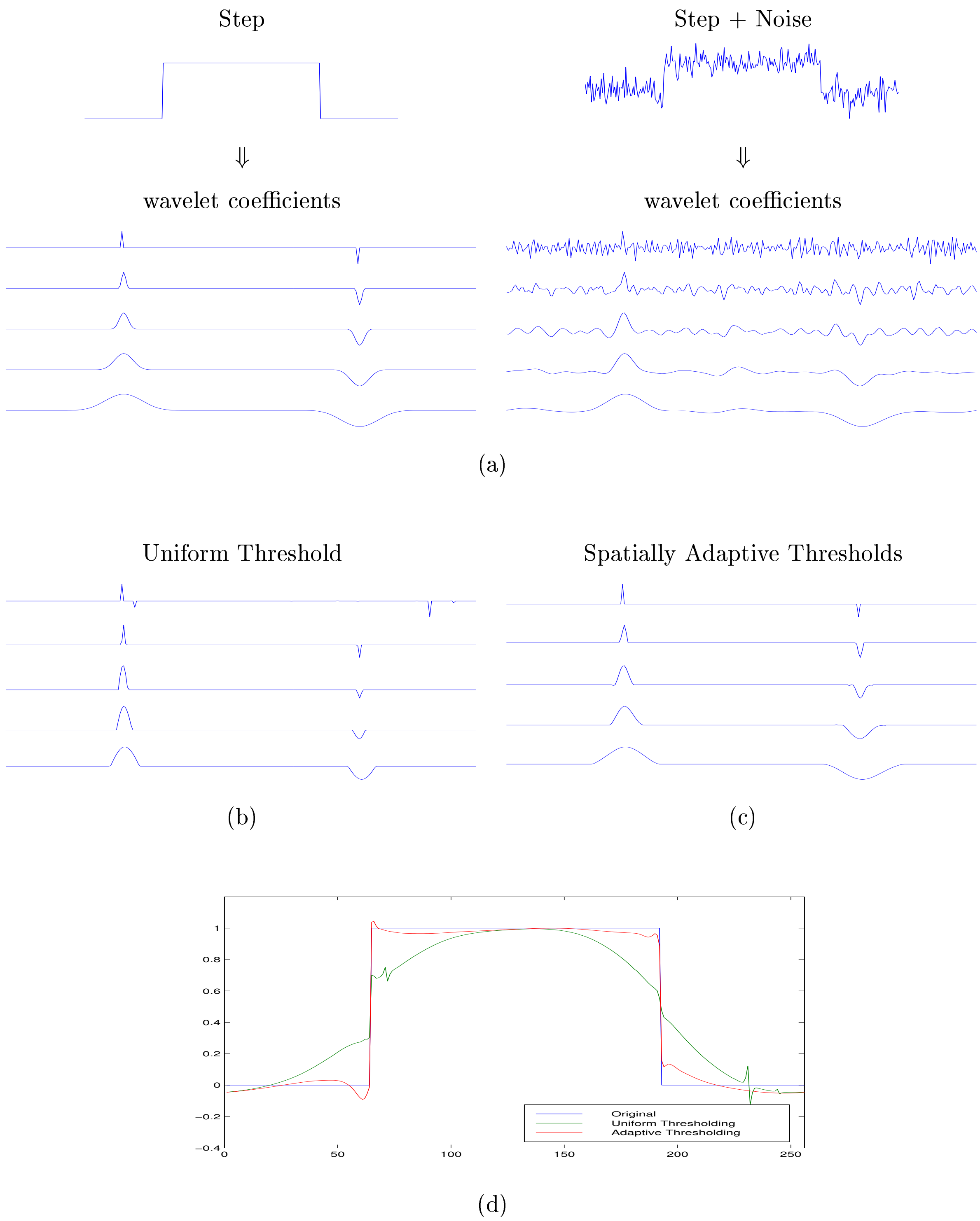

Fig. 1. Motivation for adaptive thresholds. (a) shows a step function and its noisy version, along with their wavelet decomposition of 4 scales. The wavelet coefficients are thresholded by a uniform threshold in (b) and spatially adaptive thresholds in (c). The original and the reconstructions from (b) and (c) are shown in (d).

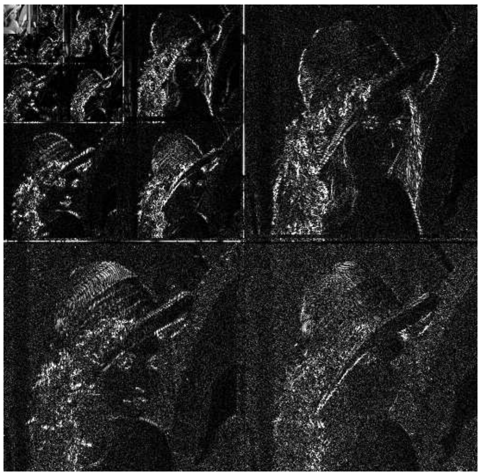

Fig. 2. Four level wavelet decomposition of *lena*. White pixels indicate large magnitude coefficients, and black signifies small magnitude.

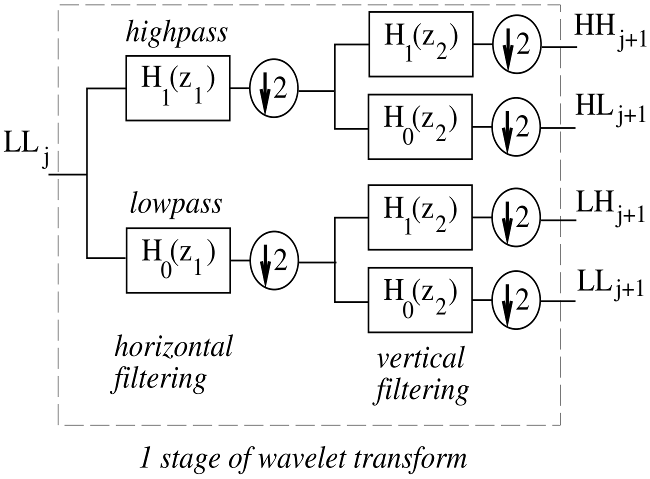

Fig. 3. One stage of the wavelet transform.

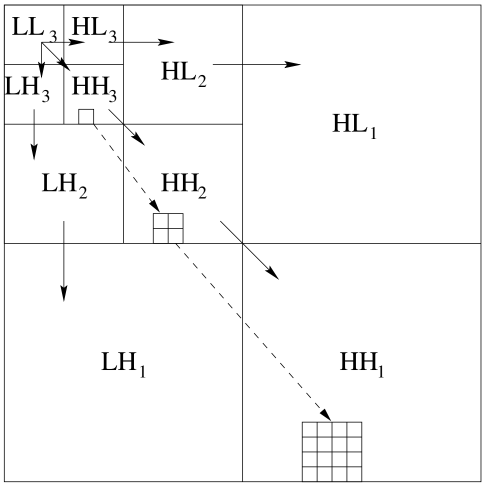

Fig. 4. Subbands of the orthogonal discrete wavelet transform(DWT), also showing the parent-child relationship.

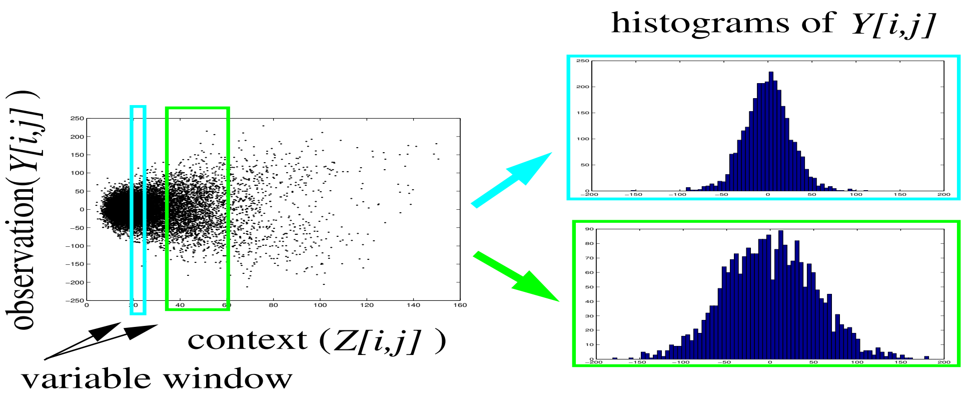

Fig. 5. A sample plot of $\{Z[i,j], Y[i,j]\}$, where $Y[i,j]$ is the noisy wavelet coefficient, and $Z[i,j]$ is its context. A collection of $Y[i,j]$ with small values of $Z[i,j]$ have a smaller spread than those with large values of $Z[i,j]$, suggesting that context modeling provides a good variability estimate of $Y[i,j]$.

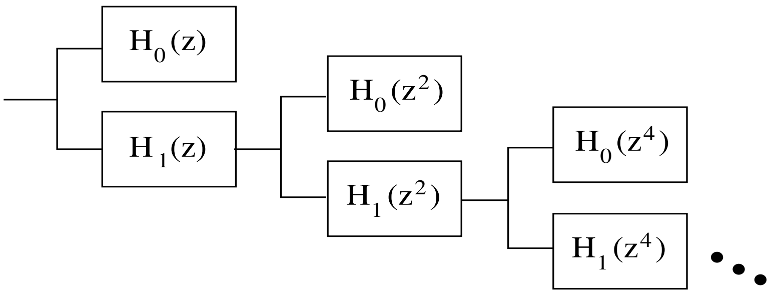

Fig. 6. The 1D non-subsampled filter bank. The 1D filtering is extended to 2D by separable filtering.

Fig. 7. Comparing results of various denoising methods, for *lena* corrupted by noise $\sigma = 22.5$ and *barbara* by noise $\sigma = 25$. Counter-clockwise from top left: original, noisy observation, adaptive thresholding in DWT basis (*AdaptDWT*), uniform thresholding in DWT basis (*OrcUnifDWT*), spatially adaptive thresholding in overcomplete expansion (*AdaptNS*), and uniform thresholding in overcomplete expansion (*OrcUnifNS*). This figure can also be found at `http://www-wavelet.eecs.berkeley.edu/~grchang/SpatialDenoise.html` .