Adaptive Wavelet Thresholding for Image Denoising and Compression

S. Grace Chang, Bin Yu, and Martin Vetterli

This work was supported in part by the NSF Graduate Fellowship and the University of California Dissertation Fellowship (Chang), in part by ARO Grant DAAH04-94-G-0232 and NSF Grant DMS-9322817 (Yu), and in part by NSF Grant MIP-93-213002 and Swiss NSF 20-52347.97 (Vetterli). S.G. Chang is formerly with the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley, and now with Hewlett-Packard Laboratories, Palo Alto, CA. B. Yu is with the Department of Statistics, University of California, Berkeley, and Bell Laboratories, Lucent Technologies, NJ. 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 Sciences at U.C. Berkeley.

Part of this work was presented at the *IEEE International Conference on Image Processing (ICIP)*, Santa Barbara, CA, USA, October 1997.

Abstract

The first part of this paper proposes an adaptive, data-driven threshold for image denoising via wavelet soft-thresholding. The threshold is derived in a Bayesian framework, and the prior used on the wavelet coefficients is the Generalized Gaussian distribution widely used in image processing applications. The proposed threshold is simple and closed-form, and it is adaptive to each subband because it depends on data-driven estimates of the parameters. Experimental results show that the proposed method, called *BayesShrink*, is typically within 5% of the MSE of the best soft-thresholding benchmark when assuming the image known. It also outperforms Donoho and Johnstone's *SureShrink* most of the time.

The second part of the paper attempts to further validate recent claims that lossy compression can be used for denoising. The *BayesShrink* threshold can aid in the parameter selection of a coder designed with the intention of denoising, and thus achieving simultaneous denoising and compression. Specifically, the zero-zone in the quantization step of compression is analogous to the threshold value in the thresholding function. The remaining coder design parameters are chosen based on a criterion derived from Rissanen's Minimum Description Length (MDL) principle. Experiments show that this compression method does indeed remove noise significantly, especially for large noise power. However, it introduces quantization noise and should be used only if bitrate were an additional concern to denoising.

I. Introduction

An image is often corrupted by noise in its acquisition or transmission. The goal of denoising is to remove the noise while retaining as much as possible the important signal features. Traditionally, this is achieved by linear processing such as Wiener filtering. A vast literature has emerged recently on signal denoising using non-linear techniques, in the setting of additive white Gaussian noise. The seminal work on signal denoising via *wavelet thresholding* or *shrinkage* of Donoho and Johnstone ([12], [13], [14], [15]) have shown that various wavelet thresholding schemes for denoising have near-optimal properties in the minimax sense and perform well in simulation studies of one-dimensional curve estimation. It has been shown to have better rates of convergence than linear methods for approximating functions in Besov spaces ([12], [13]). Thresholding is a non-linear technique, yet it is very simple because it operates on one wavelet coefficient at a time. Alternative approaches to non-linear wavelet-based denoising can be found in, for example, [1], [3], [7], [9], [8], [11], [16], [17], [22], [25], [26], [27], [30], [31], [34], and references therein.

On a seemingly unrelated front, lossy compression has been proposed for denoising in several works [4], [5], [19], [23], [26]. Concerns regarding the compression rate were explicitly addressed. This is important because any practical coder must assume a limited resource (such as bits) at its disposal for representing the data. Other works [3], [11], [12], [13], [14], [15] also addressed the connection between compression and denoising, especially with non-linear algorithms such as wavelet thresholding in a mathematical framework. However, these latter works were not concerned with quantization and bitrates: compression results from a reduced number of non-zero wavelet coefficients, and not from an explicit design of a coder.

The intuition behind using lossy compression for denoising may be explained as follows. A signal typically has structural correlations that a good coder can exploit to yield a concise representation. White noise, however, does not have structural redundancies and thus is not easily compressable. Hence, a good compression method can provide a suitable model for distinguishing between signal and noise. The discussion will be restricted to wavelet-based coders, though these insights can be extended to other transform-domain coders as well. A concrete connection between lossy compression and denoising can easily be seen when one examines the similarity between thresholding and quantization, the latter of which is a necessary step in a practical lossy coder. That is, the quantization of wavelet coefficients *with a zero-zone* is an approximation to the thresholding function (see Figure 1). Thus, provided that the quantization outside of the zero-zone does not introduce significant distortion, it follows that wavelet-based lossy compression achieves denoising. With this connection in mind, this paper is about wavelet thresholding for image denoising and also for lossy compression. The threshold choice aids the lossy coder to choose its zero-zone, and the resulting coder achieves simultaneous denoising and compression if such property is desired.

The theoretical formalization of filtering additive *iid* Gaussian noise (of zero-mean and standard deviation σ) via thresholding wavelet coefficients was pioneered by Donoho and Johnstone [13]. A wavelet coefficient is compared to a given threshold and is set to zero if its magnitude is less than the threshold; otherwise, it is kept or modified (depending on the thresholding rule). The threshold acts as an oracle which distinguishes between the insignificant coefficients likely due to noise, and the significant coefficients consisting of important signal structures. Thresholding rules are especially effective for signals with sparse or near-sparse representations where only a small subset of the coefficients represents all or most of the signal energy. Thresholding essentially creates a region around zero where the coefficients are considered negligible. Outside of this region, the thresholded coefficients are kept to full precision (that is, without quantization). Their most well-known thresholding methods include *VisuShrink* [13] and *SureShrink* [14]. These threshold choices enjoy asymptotic minimax optimalities over function spaces such as Besov spaces. For image denoising, however, *VisuShrink* is known to yield overly smoothed images. This is because its threshold choice, $\sigma\sqrt{2\log M}$ (called the *universal threshold*), can be unwarrantedly large due to its dependence on the number of samples, M, which is more than 10^5 for a typical test image of size 512×512 . *SureShrink* uses a hybrid of the universal threshold and the SURE threshold, derived from minimizing Stein's unbiased risk estimator [28], and has been shown to perform well. *SureShrink* will be the main comparison to the method proposed here, and, as will be seen later in this paper, our proposed threshold often yields better result.

Since the works of Donoho and Johnstone, there has been much research on finding thresholds for nonparametric estimation in statistics. However, few are specifically tailored for images. In this paper, we propose a framework and a near-optimal threshold in this framework more suitable for image denoising. This approach can be formally described as Bayesian, but this only describes our mathematical formulation not our philosophy. The formulation is grounded on the empirical observation that the wavelet coefficients in a subband of a natural image can be summarized adequately by a *Generalized Gaussian distribution*. This observation is well-accepted in the image compression community (for example, see [18], [20], [21], [27], [33], [35]) and is used for state-of-the-art image coders in [18], [20], [35]. It follows from this observation that the average MSE (in a subband) can be approximated by the corresponding Bayesian squared error risk with the Generalized Gaussian distribution as the prior applied to each in an *iid* fashion. That is, a sum is approximate by an integral. We emphasize that this is an analytical approximation and our framework is broader than assuming wavelet coefficients are *iid* draws from a Generalized Gaussian distribution. The goal is to find the soft-threshold that minimizes this Bayesian risk, and we call our method *BayesShrink*.

The proposed Bayesian risk minimization is subband-dependent. Given the signal being Generalized Gaussian distributed and the noise being Gaussian, via numerical calculation a nearly optimal threshold for soft-thresholding is found to be $T_B = \sigma^2/\sigma_X$ (where σ^2 is the noise variance and σ_X^2 the signal variance). This threshold gives a risk within 5% of the minimal risk over a broad range of parameters in the Generalized Gaussian distribution family. To make this threshold data-driven, the parameters σ_X and σ are estimated from the observed data, one set for each subband.

To achieve simultaneous denoising and compression, the non-zero thresholded wavelet coefficients need to be quantized. Uniform quantizer and centroid reconstruction is used on the Generalized Gaussian distribution. The design parameters of the coder, such as the number of quantization levels and binwidths, are decided based on a criterion derived from Rissanen's *Minimum Description Length* (MDL) principle [24]. This criterion balances the trade-off between the compression rate and distortion, and yields a nice interpretation of operating at a fixed slope on the rate-distortion curve.

The paper is organized as follows. In Section II, the wavelet thresholding idea is introduced. Section II-A explains the derivation of the *BayesShrink* threshold by minimizing a Bayesian risk with squared error. The lossy compression based on the MDL criterion is explained in Section III. Experimental results on several test images are shown in Section IV and compared with *SureShrink*. To benchmark against the best possible performance of a threshold estimate, the comparisons also include *OracleShrink*, the best soft-thresholding estimate obtainable assuming the original image known, and *OracleThresh*, the best hard-thresholding counterpart. The *BayesShrink* method often comes to within 5% of the MSEs of *OracleShrink*, and is better than *SureShrink* up to 8% most of the time, or is within 1% if it is worse. Furthermore, the *BayesShrink* threshold is very easy to compute. *BayesShrink* with the additional MDL-based compression, as expected, introduces quantization noise to the image. This distortion may negate the denoising achieved by thresholding, especially when σ is small. However, for larger values of σ, the MSE due to the lossy compression is still significantly lower than that of the noisy image , while fewer bits are used to code the image, thus achieving both denoising and compression.

II. Wavelet thresholding and threshold selection

Let the signal be $\{f_{ij},\ i,j = 1,\ldots,N\}$, where N is some integer power of 2. It has been corrupted by additive noise and one observes

$$g_{ij} = f_{ij} + \varepsilon_{ij}, \quad i,j = 1,\ldots,N \tag{1}$$

where $\{\varepsilon_{ij}\}$ are independent and identically distributed (*iid*) as normal $N(0,\sigma^2)$ and independent of $\{f_{ij}\}$. The goal is to remove the noise, or "denoise" $\{g_{ij}\}$, and to obtain an estimate $\{\hat{f}_{ij}\}$ of

$\{f_{ij}\}$ which minimizes the mean squared error (MSE),

$$\mathrm{MSE}(\hat{\boldsymbol{f}}) = \frac{1}{N^2} \sum_{i,j=1}^{N} (\hat{f}_{ij} - f_{ij})^2 \ . \tag{2}$$

Let $\boldsymbol{g} = \{g_{ij}\}_{i,j}$, $\boldsymbol{f} = \{f_{ij}\}_{i,j}$, $\boldsymbol{\varepsilon} = \{\varepsilon_{ij}\}_{i,j}$, that is, the boldfaced letters will denote the matrix representation of the signals under consideration. Let $\boldsymbol{Y} = \mathcal{W}\boldsymbol{g}$ denote the matrix of wavelet coefficients of $\boldsymbol{g}$, where $\mathcal{W}$ is the two-dimensional dyadic orthogonal wavelet transform operator, and similarly $\boldsymbol{X} = \mathcal{W}\boldsymbol{f}$ and $\boldsymbol{V} = \mathcal{W}\boldsymbol{\varepsilon}$. The readers are referred to references such as [21], [29] for details of the two-dimensional orthogonal wavelet transform. It is convenient to label the subbands of the transform as in Figure 2. The subbands $HH_k, HL_k, LH_k, k = 1, 2, \ldots, J$ are called the *details*, where k is the *scale*, with J being the largest (or coarsest) scale in the decomposition, and a subband at scale k has size $N/2^k \times N/2^k$. The subband LL_J is the *low resolution residual*, and J is typically chosen large enough such that $N/2^J << N$ and $N/2^J > 1$. Note that since the transform is orthogonal, $\{V_{ij}\}$ are also *iid* $N(0, \sigma^2)$.

The wavelet-thresholding denoising method filters each coefficient Y_{ij} from the detail subbands with a threshold function (to be explained shortly) to obtain $\hat{X}_{ij}$. The denoised estimate is then $\hat{\boldsymbol{f}} = \mathcal{W}^{-1}\hat{\boldsymbol{X}}$, where $\mathcal{W}^{-1}$ is the inverse wavelet transform operator.

There are two thresholding methods frequently used. The *soft-threshold* function (also called the shrinkage function)

$$\eta_T(x) = \mathrm{sgn}(x) \cdot \max(|x| - T, 0), \tag{3}$$

takes the argument and shrinks it towards zero by the *threshold* T. The other popular alternative is the *hard-threshold* function,

$$\psi_T(x) = x \cdot \mathbf{1}\{|x| > T\}, \tag{4}$$

which keeps the input if it is larger than the threshold T; otherwise, it is set to zero. The wavelet thresholding procedure removes noise by thresholding *only* the wavelet coefficients of the detail subbands, while keeping the low resolution coefficients unaltered.

The soft-thresholding rule is chosen over hard-thresholding for several reasons. First, soft-thresholding has been shown to achieve near-optimal minimax rate over a large range of Besov spaces [11], [13]. Second, for the Generalized Gaussian prior assumed in this work, the optimal soft-thresholding estimator yields a smaller risk than the optimal hard-thresholding estimator (to be shown later in this section). Lastly, in practice, the soft-thresholding method yields more visually pleasant images over hard-thresholding because the latter is discontinuous and yields abrupt artifacts in the recovered images, especially when the noise energy is significant. In what follows, soft-thresholding will be the primary focus.

While the idea of thresholding is simple and effective, finding a good threshold is not an easy task. For one-dimensional deterministic signal of length M, Donoho and Johnstone [13] proposed for *VisuShrink* the universal threshold, $T_U = \sigma\sqrt{2\log M}$, which results in an estimate asymptotically optimal in the minimax sense (minimizing the maximum error over all possible M-sample signals). One other notable threshold is the SURE threshold [14], derived from minimizing Stein's unbiased risk estimate [28] when soft-thresholding is used. The *SureShrink* method is a hybrid of the universal and the SURE threshold, with the choice being dependent on the energy of the particular subband [14]. The SURE threshold is data-driven, does not depend on M explicitly, and *SureShrink* estimates it in a subband-adaptive manner. Moreover, *SureShrink* has yielded good image denoising performance and comes close to the true minimum MSE of the optimal soft-threhsold estimator (cf. [3], [11]), and thus will be the main comparison to the proposed method.

In the statistical Bayesian literature, many works have concentrated on deriving the best threshold (or shrinkage factor) based on priors such as the Laplacian and a mixture of Gaussians (cf. [1], [7], [8], [16], [22], [25], [27], [30], [34]). With an integral approximation to the pixel-wise MSE distortion measure as discussed earlier, the formulation here is also Bayesian for finding the best soft-thresholding rule under the Generalized Gaussian prior. A related work is [25] where the hard-thresholding rule is investigated for signals with Laplacian and Gaussian distributions.

The Generalized Gaussian distribution has been used in many subband or wavelet-based image processing applications [2], [18], [20], [21], [27], [33], [35]. In [27], it was observed that a Generalized Gaussian distribution with the shape parameter β ranging from 0.5 to 1 (see Equation (5)) can adequately describe the wavelet coefficients of a large set of natural images. Our experience with the USC image database supports the same conclusion, and Figure 3 shows the histogram of the wavelet coefficients of several images, against the Generalized Gaussian curve, with the parameters labeled (the estimation of the parameters will be explained later in this text.) A heuristic can be set forward to explain why there are a large number of "small" coefficients but relatively few "large" coefficients as the Generalized Gaussian distribution suggests: the small ones correspond to smooth regions in a natural image and the large ones to edges or textures.

A. Adaptive Threshold for BayesShrink

The Generalized Gaussian distribution (GGD), following [18], is

$$GG_{\sigma_X,\beta}(x) = C(\sigma_X,\beta)\,\exp\left\{-\left[\alpha(\sigma_X,\beta)|x|\right]^\beta\right\} \quad , \quad -\infty < x < \infty, \quad \sigma_X > 0, \quad \beta > 0 \ , \tag{5}$$

where

$$\alpha(\sigma_X,\beta) = \sigma_X^{-1}\left[\frac{\Gamma(3/\beta)}{\Gamma(1/\beta)}\right]^{1/2} \quad \text{and} \quad C(\sigma_X,\beta) = \frac{\beta\cdot\alpha(\sigma_X,\beta)}{2\Gamma(\frac{1}{\beta})}$$

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 β is the *shape* parameter. For a given set of parameters, the objective is to find a soft-threshold T which minimizes the *Bayes risk*,

$$r(T) = E(\hat{X}-X)^2 = E_X E_{Y|X}(\hat{X}-X)^2 \ , \tag{6}$$

where $\hat{X} = \eta_T(Y)$, $Y|X \sim N(x,\sigma^2)$ and $X \sim GG_{\sigma_X,\beta}$. Denote the optimal threshold by T^*,

$$T^*(\sigma_X,\beta) = \arg\min_T r(T), \tag{7}$$

which is a function of the parameters σ_X and β. To our knowledge, there is no closed form solution for T^* for this chosen prior, thus numerical calculation is used to find its value[1].

Before examining the general case, it is insightful to consider two special cases of the Generalized Gaussian distribution: the Gaussian ($\beta = 2$) and the Laplacian ($\beta = 1$) distributions. The Laplacian case is particularly interesting, because it is analytically more tractable and is often used in image processing applications.

Case 1: (Gaussian) $X \sim N(0,\sigma_X^2)$ with $\beta = 2$. It is straightforward to verify that

$$\begin{aligned} E_X E_{Y|X}(\hat{X}-X)^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\eta_T(y)-x)^2 p(y|x)\,p(x)dy\,dx \\ &= \sigma^2 w\left(\frac{\sigma_X^2}{\sigma^2},\frac{T}{\sigma}\right) \end{aligned} \tag{8}$$

where

$$w(\sigma_X^2,T) = \sigma_X^2 + 2(T^2+1-\sigma_X^2)\ \ \bar{\Phi}\left(\frac{T}{\sqrt{1+\sigma_X^2}}\right) - 2T\ (1+\sigma_X^2)\ \ \phi(T,1+\sigma_X^2),$$

with the standard normal density function $\phi(x,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{x^2}{2\sigma^2})$ and the survival function of the standard normal $\bar{\Phi}(x) = \int_x^\infty \phi(t,1)\,dt$.

[1]It was observed that for the numerical calculation, it is more robust to obtain the value of T^* from locating the zero-crossing of the derivative, $r'(T)$, than from minimizing $r(T)$ directly.

Assuming $\sigma = 1$ for the time being, the value of $T^*(\sigma_X, 2)$ is found numerically for different values of σ_X and is plotted against σ_X in Figure 4(a), with

$$T_B(\sigma_X) = \frac{1}{\sigma_X} \tag{9}$$

superimposed on top. It is clear that this simple and closed-form expression, $T_B(\sigma_X)$, is very close to the numerically found $T^*(\sigma_X, 2)$. The expected risks of $T^*(\sigma_X, 2)$ and $T_B(\sigma_X)$ are shown in Figure 4(b) for $\sigma = 1$, where the maximum deviation of $r(T_B(\sigma_X))$ is less than 1% of the optimal risk, $r(T^*(\sigma_X, 2))$. For general σ, it is an easy scaling exercise to see that (9) becomes

$$T_B(\sigma_X) = \frac{\sigma^2}{\sigma_X}. \tag{10}$$

For a further comparison, the risk for hard-thresholding is also calculated. After some algebra, it can be shown that the risk for hard-thresholding is

$$r_h(T) = \sigma^2 + (\sigma^2 - \sigma_X^2)\left(2T\ \phi(T, \sigma_X^2 + \sigma^2) + 2\bar{\Phi}\left(\frac{T}{\sqrt{\sigma_X^2 + \sigma^2}}\right) - 1\right). \tag{11}$$

By setting to zero the derivative of (11) with respect to T, the optimal threshold is found to be

$$T_h^*(\sigma_X, 2) = \begin{cases} 0 & \text{if } \sigma_X > \sigma \\ \infty & \text{if } \sigma_X < \sigma \\ \text{anything} & \text{if } \sigma_X = \sigma \end{cases}, \tag{12}$$

with the associated risk

$$r_h(T_h^*) = \begin{cases} \sigma^2 & \text{if } \sigma_X > \sigma \\ \sigma_X^2 & \text{if } \sigma_X \le \sigma \end{cases}. \tag{13}$$

Figure 4(b) shows that both the optimal and near-optimal soft-threshold estimators, $\eta_{T^*}(\cdot)$ and $\eta_{T_B}(\cdot)$, achieve lower risks than the optimal hard-threshold estimator.

The threshold $T_B = \sigma^2/\sigma_X$ is not only nearly optimal but also has an intuitive appeal. The normalized threshold, T_B/σ, is inversely proportional to σ_X, the standard deviation of X, and proportional to σ, the noise standard deviation. When $\sigma/\sigma_X \ll 1$, the signal is much stronger than the noise, T_B/σ is chosen to be small in order to preserve most of the signal and remove some of the noise; vice versa, when $\sigma/\sigma_X \gg 1$, the noise dominates and the normalized threshold is chosen to be large to remove the noise which has overwhelmed the signal. Thus, this threshold choice adapts to both the signal and noise characteristics as reflected in the parameters σ and σ_X.

Case 2: (Laplacian) With $\beta = 1$, the Generalized Gaussian distribution becomes Laplacian: $X \sim \mathrm{LAP}(x) = \frac{1}{\sqrt{2}\sigma_X} \exp\{-\frac{\sqrt{2}}{\sigma_X}|x|\}$. Again for the time being let $\sigma = 1$. The optimal threshold $T^*(\sigma_X, 1)$ found numerically is plotted against the standard deviation σ_X on the horizontal axis in Figure 5(a). The curve of $T^*(\sigma_X, 1)$ (in solid line —) is compared with $T_B(\sigma_X) = 1/\sigma_X$ (in dotted line $\cdots$) in Figure 5(a). Their corresponding expected risks are shown in Figure 5(b), and the deviation of $r(T_B(\sigma_X))$ from the $r(T^*)$ is less than 0.8%. This suggests that $T_B(\sigma_X)$ also works well in the Laplacian case. For general σ, (10) holds again.

The threshold choice $T_{Bh}(\sigma_X) = 2\sqrt{2}\sigma^2/\sigma_X$ was found independently in [25] for approximating the optimal hard-thresholding using the Laplacian prior. Figure 5(a) compares the optimal hard-threshold, $T_h^*(\sigma_X, 1)$, and $T_{Bh}(\sigma_X)$ to the soft-thresholds $T^*(\sigma_X, 1)$ and $T_B(\sigma_X)$. The corresponding risks are plotted in Figure 5(b), which shows the soft-thresholding rule to yield a lower risk for this chosen prior. In fact, for σ_X larger than approximately 1.3σ, the risk of the approximate hard-threshold is worse than if no thresholding were performed (which yields a risk of σ^2).

With these insights from the special cases, the discussion now returns to the general case of GGD.

Case 3: (Generalized Gaussian) The proposed threshold $T_B(\sigma_X)$ in (10) has been found to work well for the general case. Let $\sigma = 1$. In Figure 6(a), each dotted line ($\cdots$) is the optimal threshold $T^*(\sigma_X, \beta)$ for a given fixed β, plotted against σ_X on the horizontal axis. The values $\beta = .6, 1, 2, 3, 4$ are shown. The threshold $T_B(\sigma_X) = 1/\sigma_X$ is plotted with the solid line (—). The curve of the optimal threshold that lies closest to $T_B(\sigma_X)$ is for $T^*(\sigma_X, \beta = 1)$, the Laplacian case, while other curves deviate from T_B as β moves away from 1. Figure 6(b) shows the corresponding risks. The deviation between the optimal risk $r(T^*)$ and $r(T_B)$ grows as β moves away from 1, but the error is still within 5% of the optimal $r(T^*)$ for the curves shown in Figure 6(b). Because the threshold T_B depends only on the standard deviation 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 for the wavelet coefficients typical values of β falls in the range [0.5, 1], this simple form of the threshold T_B is appropriate for our purpose. For a fixed set of parameters, the curve of the risk (as a function of the threshold T) is very flat near the optimal threshold T^*, implying that the error is not very sensitive to a slight perturbation near T^*.

B. Parameter Estimation for Data-Driven Adaptive Threshold

This section focuses on the estimation of the GGD parameters, σ_X and β, which in turn yield a data-driven estimate of $T_B(\sigma_X)$ that are adaptive to different subband characteristics.

The noise variance σ^2 needs to be estimated first. In some situations, it may be possible to measure σ^2 based on information other than the corrupted image. If such is not the case, it is estimated from the subband HH_1 by the robust median estimator used in [13], [14],

$$\hat{\sigma} = \frac{\text{Median}(|Y_{ij}|)}{.6745}, \quad Y_{ij} \in \text{ subband } HH_1. \tag{14}$$

The parameter β does not explicitly enter into the expression of $T_B(\sigma_X)$, only the signal standard deviation, σ_X, does. Therefore it suffices to estimate directly σ_X or σ_X^2.

Recall the observation model is $Y = X + V$, with X and V independent of each other, hence

$$\sigma_y^2 = \sigma_X^2 + \sigma^2 \ , \tag{15}$$

where σ_y^2 is the variance of Y. Since Y is modeled as zero-mean, σ_y^2 can be found empirically by

$$\hat{\sigma}_Y^2 = \frac{1}{n^2} \sum_{i,j=1}^{n} Y_{ij}^2, \tag{16}$$

where $n \times n$ is the size of the subband under consideration. Thus

$$\hat{T}_B(\hat{\sigma}_X) = \frac{\hat{\sigma}^2}{\hat{\sigma}_X}, \tag{17}$$

where

$$\hat{\sigma}_X = \sqrt{\max(\hat{\sigma}_Y^2 - \hat{\sigma}^2, 0)}. \tag{18}$$

In the case that $\hat{\sigma}^2 \geq \hat{\sigma}_Y^2$, $\hat{\sigma}_X$ is taken to be 0. That is, $\hat{T}_B(\hat{\sigma}_X)$ is ∞, or, in practice, $\hat{T}_B(\hat{\sigma}_X) = \max(|Y_{ij}|)$, and all coefficients are set to 0. This happens at times when σ is large (for example, $\sigma > 20$ for a grayscale image).

To summarize, we refer to our method as *BayesShrink* which performs soft-thresholding, with data-driven, subband-dependent threshold $\hat{T}_B(\hat{\sigma}_X) = \frac{\hat{\sigma}^2}{\hat{\sigma}_x}$.

III. The MDL Principle for Compression-based Denoising: The MDLQ Criterion

Recall our hypothesis is that compression achieves denoising because the zero-zone in the quantization step (typical in compression methods) corresponds to thresholding in denoising. For the purpose of compression, after using the adaptive threshold $\hat{T}_B(\hat{\sigma}_X)$ for the zero-zone, there still remains the questions of how to quantize the coefficients outside of the zero-zone and how to code them. Figure 7 illustrates the block diagram of the compression method. It shows that the coder needs to decide on the design parameters m, Δ (the number of quantization bins and binwidth, respectively), in addition to the zero-zone threshold $\hat{T}_B$. The choice of these parameters is discussed next.

When compressing a signal, two important objectives are to be kept in mind. On the one hand, the distortion between the compressed signal and the original should be kept low; on the other hand, the description of the compressed signal should use as few bits as possible to code. Typically, these two objectives are conflicting, thus a suitable criterion is needed to reach a compromise. Rissanen's *Minimum Description Length* principle allows a trade-off between these two objectives [24].

Let $\mathcal{M}$ be a library or class of models from which the "best" one is chosen to represent the data. According to the MDL principle, given a sequence of observations, the "best" model is the one that yields the shortest description length for describing the data using the model, where the description length can be interpreted as the number of bits needed for encoding. This description can be accomplished by a two-part code: one part to describe the model and the other the description of the data using the model.

More precisely, given the set of observations $\boldsymbol{Y}$, we wish to find a model $\hat{\boldsymbol{X}}$ to describe it. The MDL principle chooses $\hat{\boldsymbol{X}}$ which minimizes the two-part code-length

$$L(\boldsymbol{Y}, \hat{\boldsymbol{X}}) = L(\boldsymbol{Y}|\hat{\boldsymbol{X}}) + L(\hat{\boldsymbol{X}}), \tag{19}$$

where $L(\boldsymbol{Y}|\hat{\boldsymbol{X}})$ is the code-length for $\boldsymbol{Y}$ based on $\hat{\boldsymbol{X}}$, and $L(\hat{\boldsymbol{X}})$ is the code-length for $\hat{\boldsymbol{X}}$.

In Saito's simultaneous compression and denoising method [26] for a length-M one-dimensional signal, the hard-threshold function was used to generate the models $\hat{X} = \psi_T(Y)$, where the number K of non-zero coefficients to retain is determined by minimizing the MDL criterion. The first term $L(\boldsymbol{Y}|\hat{\boldsymbol{X}})$ is the idealized code-length with the normal distribution (see (20)), and the second term $L(\boldsymbol{X})$ is taken to be $(3/2)K\log_2 M$, of which $K\log_2 M$ are the bits needed to indicate the location of each non-zero coefficient (assuming an uniform indexing) and $(1/2)\log_2 M$ for the value of each of the K coefficients (see [24] for justification on using $(1/2)\log_2 M$ bits to store the coefficient value). Although compression has been achieved in the sense that a fewer number of nonzero coefficients are kept, [26] does not address the quantization step necessary in a practical compression setting.

In the following section, an MDL-based quantization criterion will be developed by minimizing $L(\boldsymbol{Y}, \hat{\boldsymbol{X}})$ with the restriction that $\hat{\boldsymbol{X}}$ belongs to the set of quantized signals.

A. Derivation of the MDLQ Criterion

Consider a particular subband of size $n \times n$. Since the noisy wavelet transform coefficients are $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{V}$, where V_{ij} are *iid* $N(0, \sigma^2)$, then $Y_{ij}|X_{ij} \sim N(X_{ij}, \sigma^2)$. Thus,

$$\begin{aligned} L(\boldsymbol{Y}|\boldsymbol{X}) &= -\sum_{i,j=1}^{n} \log p(Y_{ij}|X_{ij}) \\ &= \frac{1}{2\sigma^2 \log 2} \sum_{i,j=1}^{n} (Y_{ij} - X_{ij})^2 + \frac{1}{2} \log(2\pi\sigma^{2n^2}). \end{aligned} \tag{20}$$

The second term in (20) is a constant, and thus is ignored in the minimization. Expression (20) appears also in [19], [26]. The approach described here differs from theirs in the estimate $\boldsymbol{X}$.

Let $\mathcal{M}$ be the set of quantized coefficients, $\hat{\boldsymbol{X}}^Q$, and $\hat{\boldsymbol{X}}$ be constrained in $\mathcal{M}$, then (20) (with constant terms removed) becomes

$$\begin{aligned} L(\boldsymbol{Y}|\hat{\boldsymbol{X}}^Q) &= -\sum_{i,j=1}^{n} \log p(Y_{ij}|\hat{X}^Q_{ij}) \\ &= \frac{1}{2\sigma^2 \log 2} \sum_{i,j=1}^{n} (Y_{ij} - \hat{X}^Q_{ij})^2. \end{aligned} \tag{21}$$

There are many possible ways to quantize and encode $\hat{X}$. One way is the *uniform threshold quantizer* (UTQ) with centroid reconstruction based on the Generalized Gaussian distribution. The parameters of the GGD can be estimated from the observed noisy coefficients as described below, which is a variant of that described in [27].

For noiseless observations, σ_X^2 is estimated as

$$\hat{\sigma}_x^2 = \frac{1}{n^2} \sum_{i,j=1}^{n} X_{ij}^2 \tag{22}$$

and β is solved from

$$\kappa_X = \frac{\Gamma(\frac{1}{\beta})\Gamma(\frac{5}{\beta})}{\Gamma^2(\frac{3}{\beta})} , \tag{23}$$

where κ_X is the kurtosis of the GGD and is estimated as $\hat{\kappa}_X = \dfrac{1}{\hat{\sigma}_X^4} \dfrac{1}{n^2} \sum_{i,j=1}^{n} X_{ij}^4$. The parameter values listed in Figure 3 are estimated this way.

When the image is corrupted by additive Gaussian noise, the second and fourth moments have the following relations:

$$\begin{aligned} \sigma_Y^2 &= \sigma_X^2 + \sigma^2 \ , \\ \kappa_Y &= \frac{1}{\sigma_Y^4}\left(6\sigma^2\sigma_Y^2 - 3\sigma^4 + (\sigma_Y^2 - \sigma^2)^2 \ \frac{\Gamma(\frac{1}{\beta})\Gamma(\frac{5}{\beta})}{\Gamma^2(\frac{3}{\beta})}\right) \ . \end{aligned} \tag{24}$$

The noise variance, σ^2, is estimated via (14). The second moment, σ_y^2, and the kurtosis, κ_Y, can be measured from the observations $\{Y_{ij}\}$. The parameter σ_X is estimated as in (18) and β is then solved from (24).

Once the GGD parameters have been estimated, the quantizer has sufficient information to perform the quantization. The quantizer, shown in Figure 8, consists of m levels of bins of equal size Δ on each side, resulting in a total of $2m+1$ quantization bins (one zero-zone plus m symmetric levels on each positive and negative side). These bins are indexed as $\ell = -m, \ldots, -1, 0, 1, \ldots, m$. Consider the positive side and let $b_0, b_1, \ldots, b_m$ denote the boundaries of the quantization bins, with centroid reconstruction values $\gamma_1, \gamma_2, \ldots, \gamma_m$. The value of γ_ℓ with boundaries $b_{\ell-1}$ and b_ℓ is

$$\gamma_\ell = \frac{\int_{b_{\ell-1}}^{b_\ell} x \ GG_{\sigma_X,\beta}(x)dx}{\int_{b_{\ell-1}}^{b_\ell} GG_{\sigma_X,\beta}(x)dx} \ . \tag{25}$$

Equation (25) is calculated using numerical integration (e.g. the trapezoidal rule) since it does not have a closed-form solution. Note that $b_0 = 0$, and during quantization, b_m is taken to be ∞.

The negative side is quantized in a symmetric way. The quantized coefficients are denoted by $\{\hat{X}_{ij}^Q\}$. Note that the zero coefficients resulting from thresholding are kept as zeros, and that the subsequent quantization of the non-zero coefficients does not set any additional coefficients to zero. On average, the smallest number of bits needed to code $\hat{\boldsymbol{X}}^Q$ is the Shannon code. Thus the code-length for coding the bin indices is

$$L(\hat{\boldsymbol{X}}^Q|m,\Delta) = -\sum_{\ell=-m}^{m} K_\ell \ \log\frac{K_\ell}{n^2} \ , \tag{26}$$

where K_ℓ is the number of coefficients in bin ℓ. The additional parameters m and Δ need to be coded also, but it is supposed that any positive values are equally likely, thus a fixed number of bits are allocated for $L(m,\Delta)$ for each subband (8 bytes were used in the experiment).

Now we state the model selection criterion, *MDLQ*:

$$MDLQ(\boldsymbol{X}^Q, m, \Delta) = \frac{1}{2\sigma^2 \log 2} \sum_{i,j=1}^{n} (Y_{ij} - \hat{X}_{ij}^Q)^2 + L(\hat{\boldsymbol{X}}^Q | m, \Delta) \ , \tag{27}$$

$$\text{with} \quad L(\hat{\boldsymbol{X}}^Q | m, \Delta) = - \sum_{\ell=-m}^{m} K_\ell \log \frac{K_\ell}{n^2} \ .$$

To find the best model, (27) is minimized over m and Δ to find the corresponding set of quantized coefficients. In the implementation, which is by no means optimized, this is done by searching over $m = 1, 2, \ldots,$ and for each m, minimizing (27) over a reasonable range of Δ (such as $\Delta \in [\max(|Y_{ij}|)/(m+5), \max(|Y_{ij}|)/(m-1.5)]$). Typically, once a minimum (in m) has been reached, the *MDLQ* cost increase monotonically, thus the search can terminate soon after a minimum has been detected.

This *MDLQ* compression with *BayesShrink* zero-zone selection is applied to each subband independently. The steps discussed in this section are summarized here:

- Estimate the noise variance σ^2, and the GGD standard deviation σ_X.
- Calculate the threshold $\hat{T}_B$, and soft-threshold the wavelet coefficients.
- To quantize the non-zero coefficients, minimize (27) over m and Δ to find the corresponding quantized coefficients $\hat{\boldsymbol{X}}^Q$, which is the compressed, denoised estimate of $\boldsymbol{X}$.

The coarsest subband LL_J is quantized differently in that it is not thresholded, and the quantization with (27) assumes the uniform distribution. The LL_J coefficients are essentially local averages of the image, and are not characterized by distributions with a peak at zero, thus the uniform distribution is used for generality. With the mean subtracted, the uniform distribution is assumed to be symmetric about zero. Every quantization bin (including the zero-zone) is of width Δ, and the reconstruction values are the midpoints of the intervals.

The *MDLQ* criterion in (27) has the additional interpretation of operating at a specified point on the rate-distortion (R-D) curve, as also pointed out by Liu and Moulin [19]. For a given coder, one can obtain a set of operational rate-distortion points (R, D). When there is a rate or distortion constraint, the constraint problem can be formulated into a minimization problem with a Lagrange multiplier, $\lambda D + R$. In this case, (27) can be interpreted as operating at $\lambda = \frac{1}{2\sigma^2 \log 2}$. Natarajan [23] and Liu and Moulin [19] both proposed to use compression for denoising. The former operates at a constrained distortion, $D \leq \sigma^2$, and the latter operates at $\lambda = \frac{1}{2\sigma^2 \log 2}$ on the R-D curve. Both works recommend the use of "any reasonable coder" while our coder is designed specifically with the purpose of denoising.

IV. Experimental Results and Discussion

The 512×512 grayscale images "*goldhill*", "*lena*", "*barbara*" and "*baboon*" are used as test images with different noise levels $\sigma = 10, 20, 30, 35$. The original images are shown in Figure 9. The wavelet transform employs Daubechies' least asymmetric compactly-supported wavelet with 8 vanishing moments [10] with 4 scales of orthogonal decomposition.

To assess the performance of *BayesShrink*, it is compared with *SureShrink* [14]. The 1-D implementation of *SureShrink* can be obtained from the WaveLab toolkit [32], and the 2-D extension is straightforward. To gauge the best possible performance of a soft-threshold estimator, these methods are also benchmarked against what we call *OracleShrink*, which is the truly optimal soft-thresholding estimator assuming the original image is known. The threshold of *OracleShrink* in each subband is

$$T_{OS} = \arg\min_T \sum_{i,j=1}^{n} (\eta_T(Y_{ij}) - X_{ij})^2 \tag{28}$$

with X_{ij} known. To further justify the choice of soft-thresholding over hard-thresholding, another benchmark, *OracleThresh*, is also computed. *OracleThresh* is the best possible performance of a hard-threshold estimator, with subband-adaptive thresholds, each of which is defined as

$$T_{OT} = \arg\min_T \sum_{i,j=1}^{n} (\psi_T(Y_{ij}) - X_{ij})^2 \tag{29}$$

with X_{ij} known. The MSEs from the various methods are compared in Table I, and the data are collected from an average of 5 runs. The columns refer to, respectively, *OracleShrink*, *SureShrink*, *BayesShrink*, *BayesShrink* with additional MDLQ-based compression, *OracleThresh*, and the bitrate (in bpp, or bits-per-pixel) of the MDLQ-compressed image. Since the main benchmark is against *SureShrink*, the better one of the *SureShrink* and *BayesShrink* is highlighted in bold font for each test set. The MSEs resulting from *BayesShrink* comes to within 5% of *OracleShrink* for the smoother images *goldhill* and *lena*, and are most of the time within 6% for highly detailed images such as *barbara* and *baboon* (though it may suffer up to 20% for small σ). *BayesShrink* outperforms *SureShrink* most of the time, up to approximately 8%. We observed in the experiments that using solely the SURE threshold yields excellent performance (sometimes yielding even lower MSE than *BayesShrink* by up to 1-2%). However, the hybrid method of *SureShrink* results at times in the choice of the universal threshold which can be too large. As illustrated in Table I, all three soft-thresholding methods outperforms significantly the best hard-thresholding rule, *OracleThresh*.

It is not surprising that the SURE threshold and the *BayesShrink* threshold yield similar performances. The SURE threshold can also be viewed as an approximate optimal soft-threshold in terms of MSE. For a particular subband of size $n \times n$, following [14],

$$Sure(T, \boldsymbol{Y}) = n^2 - 2 \sum_{i,j=1}^{n} I_{\{i:|Y_{ij}| \le T\}} + \sum_{i,j=1}^{n} (|Y_{ij}| \wedge T)^2,$$

where $a \wedge b$ denotes $\min(a, b)$, and the SURE threshold is defined to be the value of T minimizing $Sure(T, \boldsymbol{Y})$.

Recall

$$Y_{ij} = X_{ij} + V_{ij}, \quad i, j = 1, \ldots, n$$

and V_{ij}'s are *iid* $N(0, \sigma^2)$. Conditioning on $\boldsymbol{X} = \boldsymbol{x}$, by Stein's result,

$$E_{\boldsymbol{Y}}[||\eta_T(\boldsymbol{Y}) - \boldsymbol{x}||^2 | \boldsymbol{X} = \boldsymbol{x}] = E_{\boldsymbol{Y}}[Sure(T, \boldsymbol{Y}) | \boldsymbol{X} = \boldsymbol{x}] \; . \tag{30}$$

Moreover, as we have done before, if the distribution of X's is approximated by a GGD, then the distribution of Y's is approximated by the mixture distribution of GGD and $N(0, \sigma^2)$; or $Y = X + V$ while X follows a GGD and independent of V.

By the Law of Large Numbers,

$$\frac{1}{n^2} Sure(T, \boldsymbol{Y}) \approx 1 - 2E_Y I_{\{|Y| \le T\}} + E_Y(|Y| \wedge T)^2. \tag{31}$$

Taking expectation with respect to the GGD on both sides of (30), the risk can be written as

$$\begin{aligned} r(T) &= \frac{1}{n^2} E_Y ||\eta_T(Y) - x||^2 \\ &= \frac{1}{n^2} E_{\boldsymbol{Y}} Sure(T, \boldsymbol{Y}) \\ &= 1 - 2E_Y I_{\{|Y| \le T\}} + E_Y(|Y| \wedge T)^2. \end{aligned} \tag{32}$$

Comparing (32) with (31), one can conclude that $\frac{1}{n^2} Sure(T, \boldsymbol{Y})$ is a data-based approxmation to $r(T)$, and the SURE threshold, which minimizes $Sure(T, \boldsymbol{Y})$, is an alternative to *BayesShrink* for minimizing the Bayesian risk.

The MDLQ-based compression step introduces quantization noise which is quite visible. As shown in the last column of Table I, the coder achieves a lower bitrate, but at the expense of increasing the MSE. The MSE can be even worse than the noisy observation for small values of σ, especially for the highly detailed images. This is because the quantization noise is significant compared to the additive Gaussian noise. For larger σ, the compressed images can achieve noise reduction up to approximately 75% in terms of MSE. Furthermore, the bitrates are significantly less than the original 8 bpp for grayscale images. Thus, compression does achieve denoising and the proposed MDLQ-based compression can be used if simultaneous denoising and compression is a desired feature. If only the best denoising performance were the goal, obviously using solely *BayesShrink* is preferred.

Note that the first-order entropy coding, $L(\hat{\boldsymbol{X}}^Q|m, \Delta)$, for the bitrate of the quantized coefficients is a rather loose estimate. With more sophisticated coding methods (e.g. predictive coding, pixel classification), the same bitrate could yield a higher number of quantization level m, thus resulting in a lower MSE.

Figure 10 shows the resulting images of each denoising method for *goldhill* and $\sigma = 20$. Table II compares the threshold values for each subband chosen by *OracleShrink, SureShrink* and *BayesShrink*, averaged over 5 runs. It is clear that the *BayesShrink* threshold selection is comparable to the SURE threshold and to the true optimal threshold T_{OS}. Some of the unexpectedly large threshold values in *SureShrink* comes from the universal threshold, not the SURE threshold, and these are placed in parentheses in the table. Table II (c) lists the thresholds of *BayesShrink*, and the thresholds in parentheses correspond to the case when $\hat{T}_B = \max(|Y_{ij}|)$, and all coefficients have been set to zero. Table III tabulates the values of m chosen by $MDLQ$ for each subband of the *goldhill* image, $\sigma = 20$, averaged over 5 runs. The $MDLQ$ criterion allocates more levels in the coarser, more important levels, as would be the case in a practical subband coding situation. A value of $m = 0$ indicates that the coefficients have already been thresholded to zero, and there is nothing to code.

The results for *lena* and $\sigma = 10$ are also shown. Figure 11 shows the same sequences for a zoomed-in portion of *lena* with noise $\sigma = 10$ The corresponding results of threshold selections and $MDLQ$ parameters for *lena* with noise $\sigma = 10$ are listed in Tables IV and V. Interested readers can obtain a better view of the images at the website,
`http://www-wavelet.eecs.berkeley.edu/~grchang/compressDenoise/` .

V. Conclusion

Two main issues regarding image denoising were addressed in this paper. Firstly, an adaptive threshold for wavelet thresholding images was proposed, based on the Generalized Gaussian distribution modeling of subband coefficients, and test results showed excellent performance. Secondly, a coder was designed specifically for simultaneous compression and denoising. The proposed *BayesShrink* threshold specifies the zero-zone of the quantization step of this coder, and this zero-zone is the main agent in the coder which removes the noise. Although the setting in this paper was in the wavelet domain, the idea can be extended to other transform domains such as DCT, which also relies on the energy compaction and sparse representation properties to achieve good compression.

There are several interesting directions worth pursuing. The current compression selects the threshold (i.e. zero-zone size) T_B and the quantization bin size Δ in a two-stage process. In typical image coders, however, the zero-zone is chosen to be either the same size or twice the size as other bins. Thus it would be interesting to jointly select these two values and analyze their dependencies on each other. Furthermore, a more sophisticated coder is likely to produce better compressed images than the current scheme, which uses the first order entropy to code the bin indices. With an improved coder, an increase in the number of quantization bins would not increase the bitrate penalty by much, and thus the coefficients would be quantized at a finer resolution than the current method. Lastly, the model family $\mathcal{M}$ could be expanded. For example, one could use a collection of wavelet bases for the wavelet decomposition, rather than using just one chosen wavelet, to allow possibly better representations of the signals.

In our other work [6], it was demonstrated that *spatially* adaptive thresholds greatly improves the denoising performance over uniform thresholds. That is, the threshold value changes for *each* coefficient. The threshold selection uses the context-modeling idea prevelant in coding methods, thus it would be interesting to extend this spatially adaptive threshold to the compression framework, without incuring too much overhead. This would likely improve the denoising performance.

References

[1] F. Abramovich, T. Sapatinas, and B. Silverman, "Wavelet Thresholding via a Bayesian Approach," preprint, 1996.

[2] M. Antonini, M. Barlaud, P. Mathieu, I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. Image Processing*, vol. 1, no. 2, pp. 205-220, 1992.

[3] A. Chambolle, R. A. DeVore, N. Lee, and B.J. Lucier, "Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. Image Processing*, vol 7, pp. 319-335, 1998.

[4] S.G. Chang, B. Yu, and M. Vetterli, "Bridging compression to wavelet thresholding as a denoising method," *Proc. Conf. Information Sciences and Systems*, pp. 568-573, Baltimore, MD, March 1997.

[5] 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, Santa Barbara, CA, November 1997.

[6] S.G. Chang, B. Yu, and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," in *Proc. IEEE Int. Conf. Image Processing*, Chicago, IL, October 1998.

[7] H. Chipman, E. Kolaczyk, and R. McCulloch, "Adaptive bayesian wavelet shrinkage," *J. Amer. Statist. Assoc.* vol. 92, no. 440, pp. 1413-1421, 1997.

[8] M. Clyde, G. Parmigiani, and B. Vidakovic, "Multiple shrinkage and subset selection in wavelets," *Discussion Paper* **95-37**, ISDS, Duke University, Durham, NC.

[9] M.S. Crouse, R.D. Nowak, R.G. Baraniuk. "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 46, no. 4, pp. 886-902, 1998.

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

[11] R.A. DeVore and B.J. Lucier, "Fast wavelet techniques for near-optimal image processing," *IEEE Military Communications Conference Record*, San Diego, Oct. 11-14, IEEE, Piscataway, NJ, pp. 1129-1135, 1992.

[12] D.L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Information Theory*, vol. 41, no. 3, pp. 613-627, May 1995.

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

[14] D.L. Donoho and I.M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Assoc.*, vol. 90, no. 432, pp. 1200-1224, December 1995.

[15] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet Shrinkage: Asymptopia?" *Journal of the Royal Statistical Society*, Ser. B, **57**, No. 2, pp. 301-369, 1995.

[16] M. Jansen, M. Malfait, and A. Bultheel, "Generalized cross validation for wavelet thresholding," preprint, Dec. 1995.

[17] 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.

[18] R.L. Joshi, V.J. Crump, and T.R. Fisher, "Image subband coding using arithmetic and trellis coded quantization," submitted to *IEEE Trans. Circuits and Systems for Video Technology*, 1994.

[19] J. Liu and P. Moulin, "Complexity-regularized image denoising," *Proc. IEEE Int. Conf. Image Processing*, vol. II, pp. 370-373, October 1997.

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

[21] 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.

[22] G. Nason, "Choice of the threshold parameter in wavelet function estimation," in *Wavelets in Statistics*, Eds. A. Antoniades and G. Oppenheim. Lecture Notes in Statistics, Springer-Verlag, 1995.

[23] B.K. Natarajan, "Filtering random noise from deterministic signals via data compression," *IEEE Trans. on Signal Processing*, vol. 43, no. 11, pp. 2595-2605, November 1995.

[24] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, 1989.

[25] F. Ruggeri and B. Vidakovic, "A Bayesian decision theoretic approach to wavelet thresholding," *J. Amer. Statist. Assoc.* vol. 93(441), pp. 173-179, 1998.

[26] N. Saito, "Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion," in *Wavelets in Geophysics*, Eds. E. Foufoula-Georgiou and P. Kumar, pp. 299-324, Academic Press, 1994.

[27] E. Simoncelli and E. Adelson, "Noise removal via Bayesian wavelet coring," In *Proc. IEEE Int. Conf. Image Processing, J. Amer. Statist. Assoc.* vol. I, pp. 379-382, September 1996.

[28] C.M. Stein, "Estimation of the mean of a multivariate normal distribution," *Annals of Statistics*, vol. 9, no. 6, pp. 1135-1151, 1981.

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

[30] B. Vidakovic, "Nonlinear wavelet shrinkage with Bayes rules and Bayes factors," Discussion paper 94-24, ISDS, Duke University, Durham, NC, 1994. (ftp://ftp.isds.duke.edu/pub/Users/brani/papers/WavShrinkBF.ps.Z).

[31] Y. Wang, "Function estimation via Wavelets for data with long-range dependence," *Technical Report*, Dept. of Statistics, University of Missouri-Columbia, 1994.

[32] J. Buckheit, S. Chen, D. Donoho, I. Johnstone, and J. Scargle, *WaveLab Toolkit*, `http://www-stat.stanford.edu:80/~wavelab/`.

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

[34] N. Weyrich and G.T. Warhola, "De-noising using wavelets and cross-validation," *Technical Report* AFIT/EN/TR/94-01, Dept. of Mathematics and Statistics, Air Force Inst. of Tech., AFIT/ENC, Ohio, 1994.

[35] Y. Yoo, A. Ortega, and B. Yu, "Image subband coding using context based classification and adaptive quantization," submitted to *IEEE Trans. Image Processing* (revised), 1997.

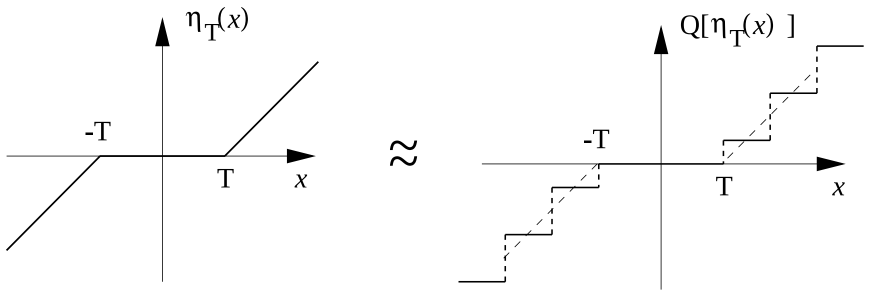


Fig. 1. The thresholding function can be approximated by quantization with a zero-zone.

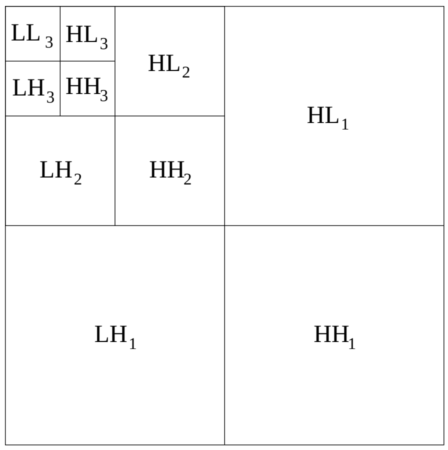
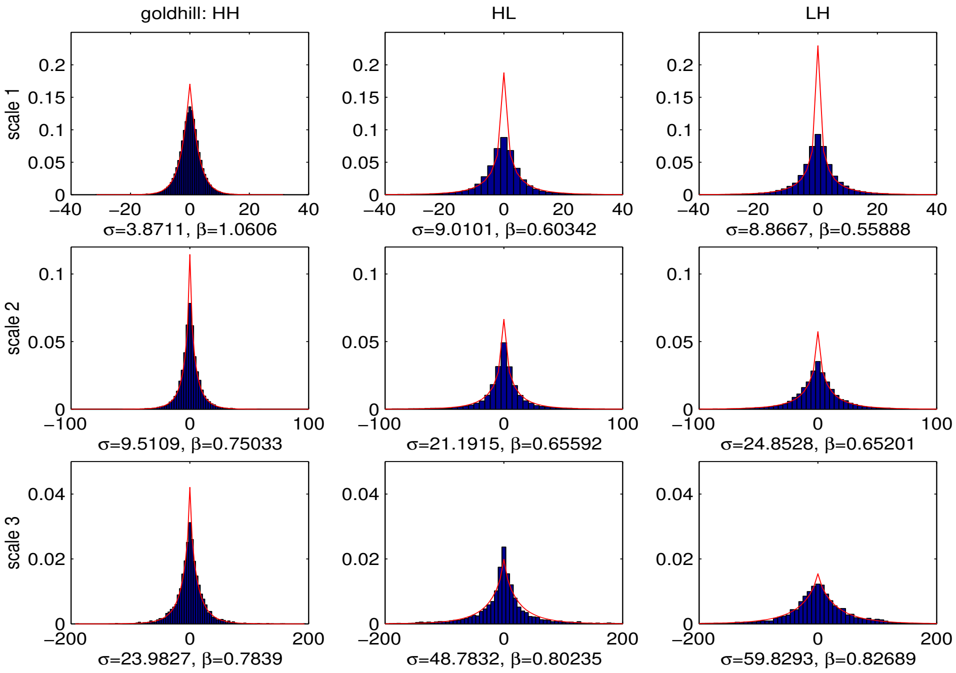
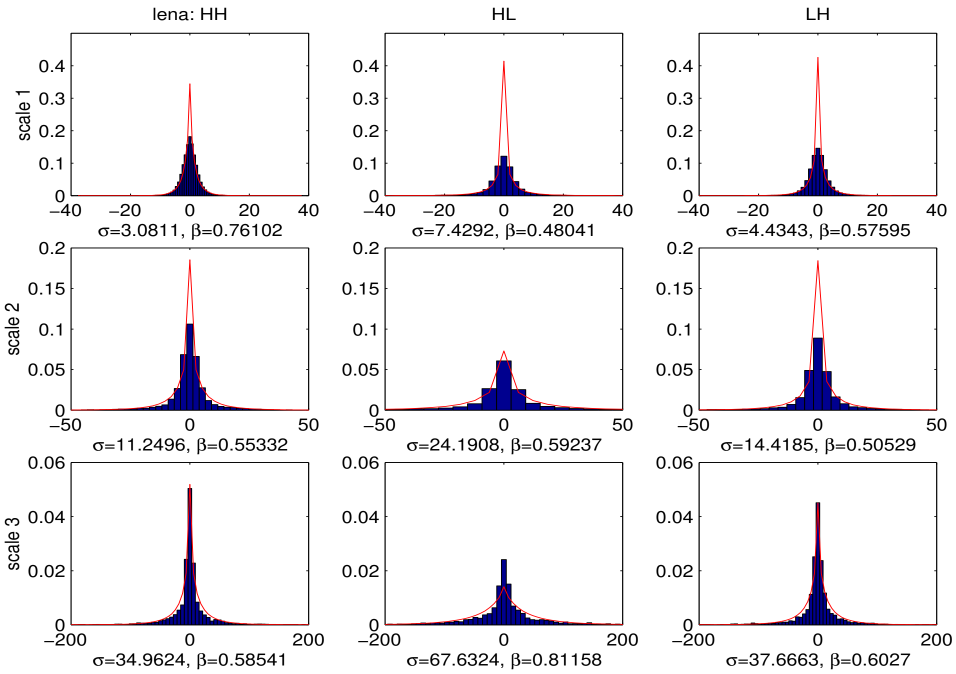


Fig. 2. Subbands of the 2D orthogonal wavelet transform.

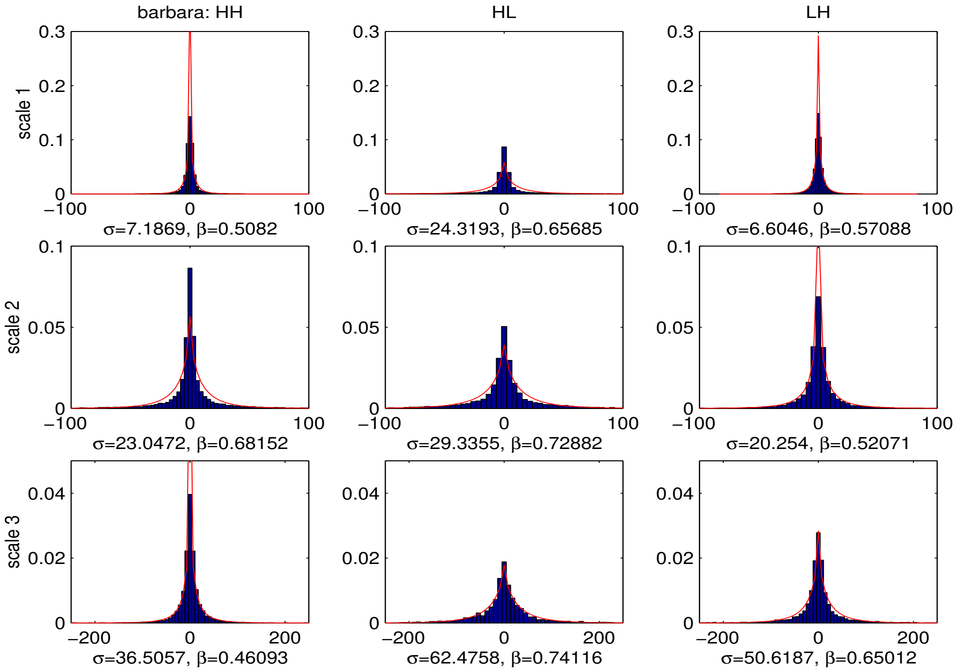


(a) *goldhill*

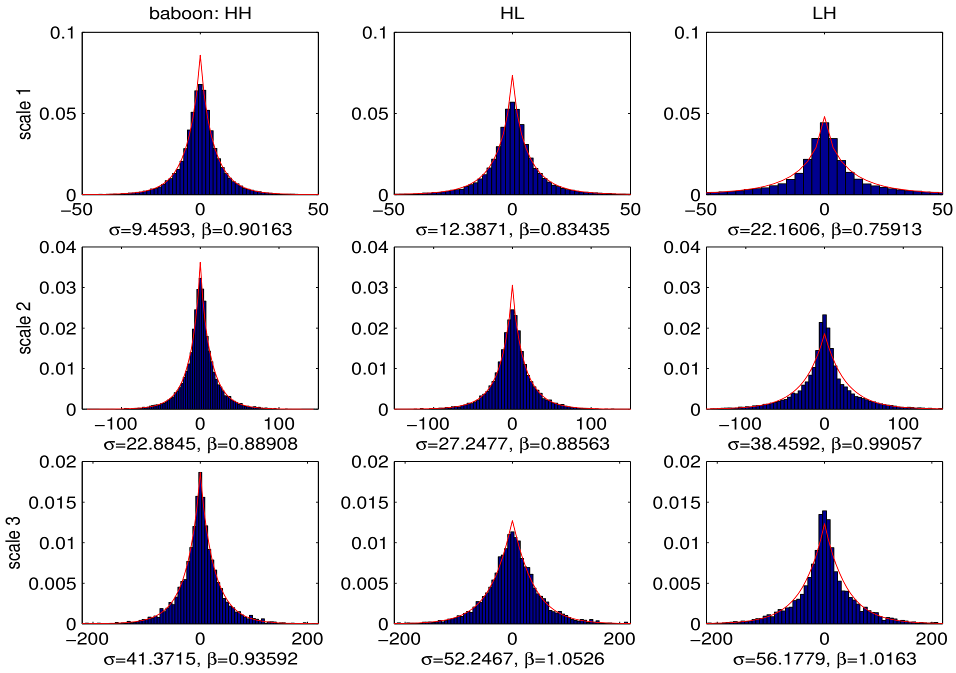


(b) *lena*

Fig. 3. Histogram of the wavelet coefficients. For each image, from top to bottom it is fine to coarse scales; from left to right, they are the *HH*, *HL*, and *LH* subbands, respectively.

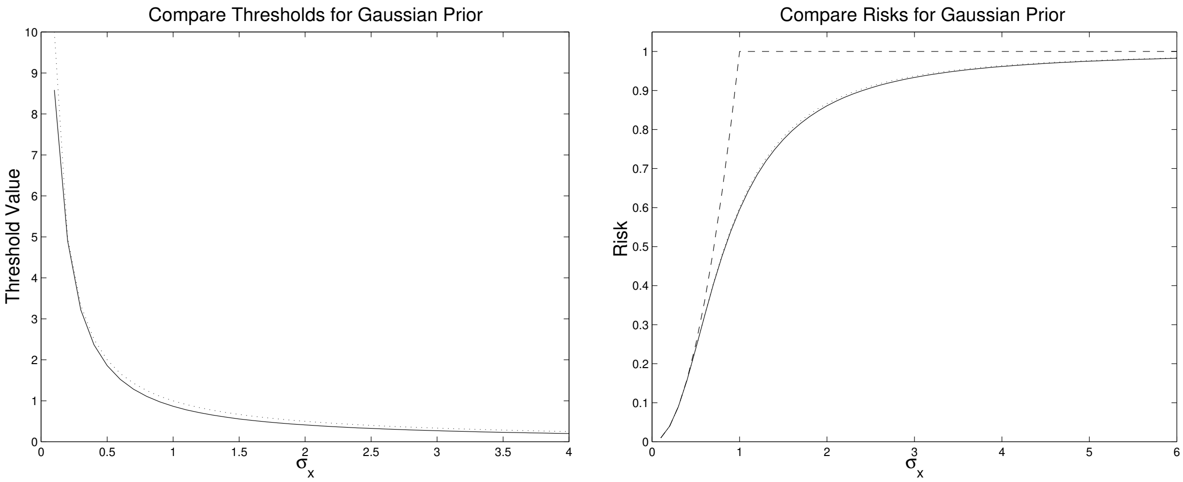


(c) *barbara*



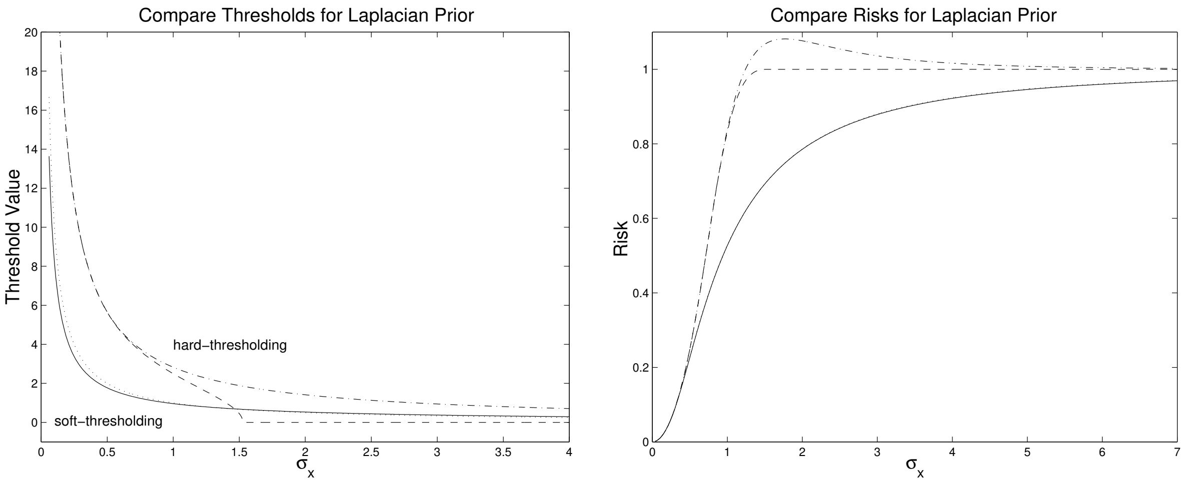
(d) *baboon*

Fig. 3. Continued.



(a) (b)

Fig. 4. Thresholding for the Gaussian prior, with $\sigma = 1$. (a) Compare the optimal threshold $T^*(\sigma_X, 2)$ (solid —) and the threshold $T_B(\sigma_X)$ (dotted $\cdots$) against the standard deviation σ_X on the horizontal axis. (b) Compare the risk of using optimal soft-thresholding (—), T_B for soft-thresholding ($\cdots$), and optimal hard-thresholding ($- - -$).



(a) (b)

Fig. 5. Thresholding for the Laplacian prior, with $\sigma = 1$. (a) Compare the optimal soft-threshold $T^*(\sigma_X, 1)$ (—), the *BayesShrink* threshold $T_B(\sigma_X)$ ($\cdots$), the optimal hard-threshold $T_h^*(\sigma_X, 1)$ ($- - -$), and the threshold T_{Bh} ($- \cdot -\cdot$) against the standard deviation on the horizontal axis. (b) Their corresponding risks.

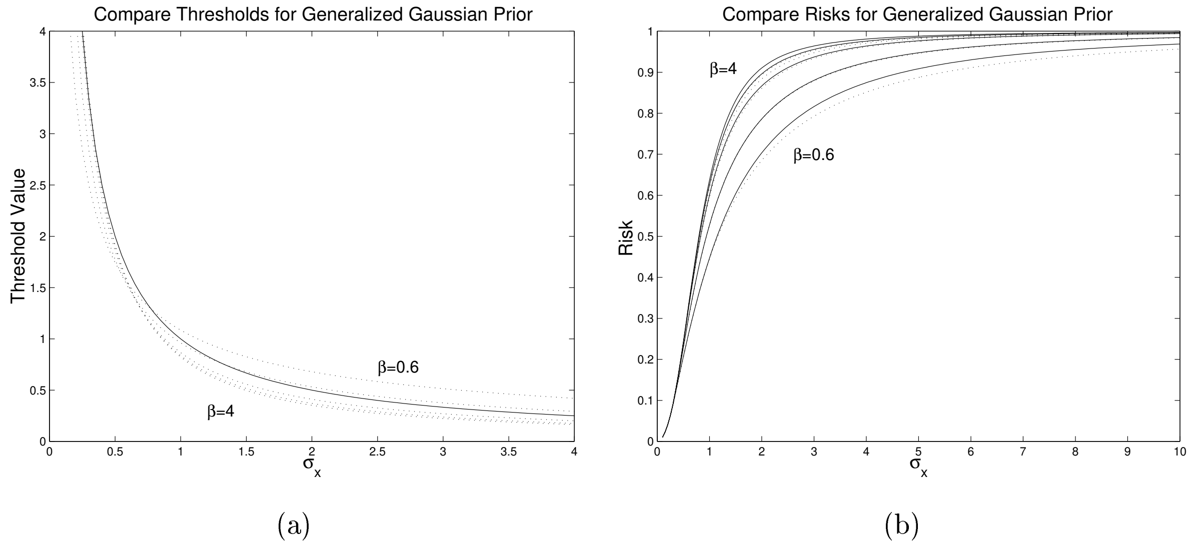


Fig. 6. Thresholding for the Generalized Gaussian prior, with $\sigma = 1$. (a) Compare the approximation $T_B(\sigma_X) = \sigma^2/\sigma_X$ (—) with the optimal threshold $T^*(\sigma_X, \beta)$ for $\beta = 0.6, 1, 2, 3, 4$ $(\cdots)$. The horizontal axis is the standard deviation, σ_X. (b) The optimal risks are in $(\cdots)$, and the approximation in (—).

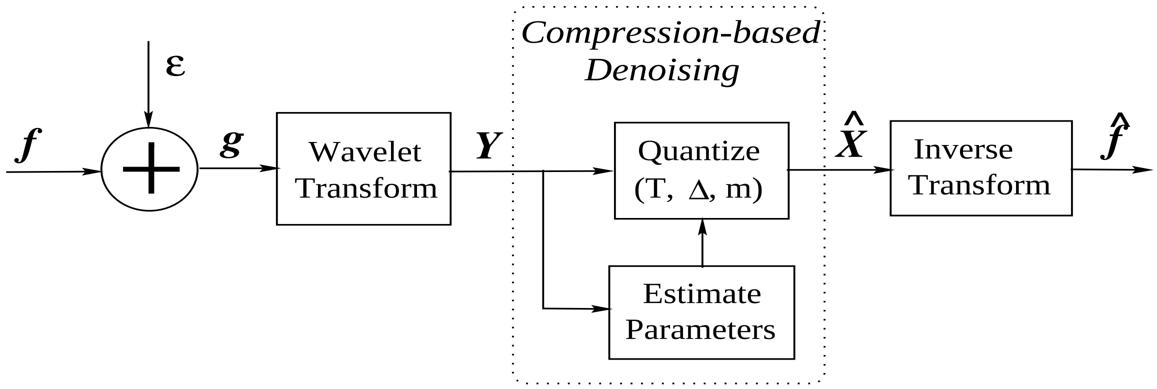


Fig. 7. Schematic for compression-based denoising. Denoising is achieved in the wavelet transform domain by lossy-compression, which involves the design of parameters T, m, and Δ, relating to the zero-zone width, the number of quantization levels, and the quantization binwidth, respectively.

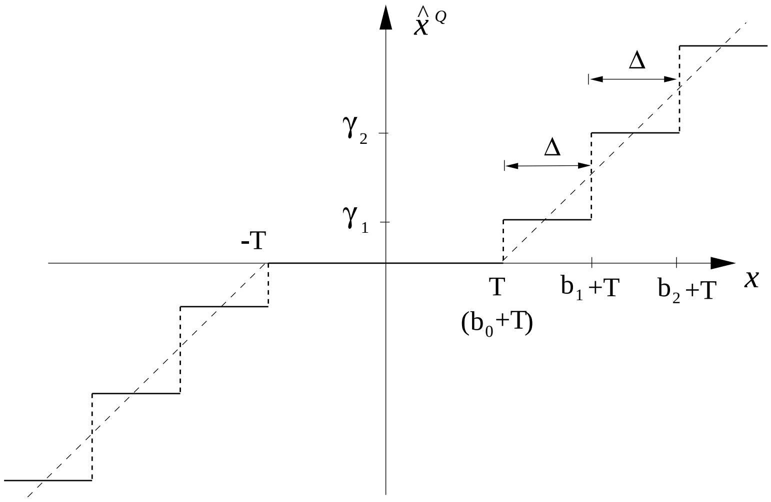


Fig. 8. Illustrating the quantizer.

Fig. 9. Original images. From top left, clockwise: *goldhill, lena, barbara* and *baboon.*

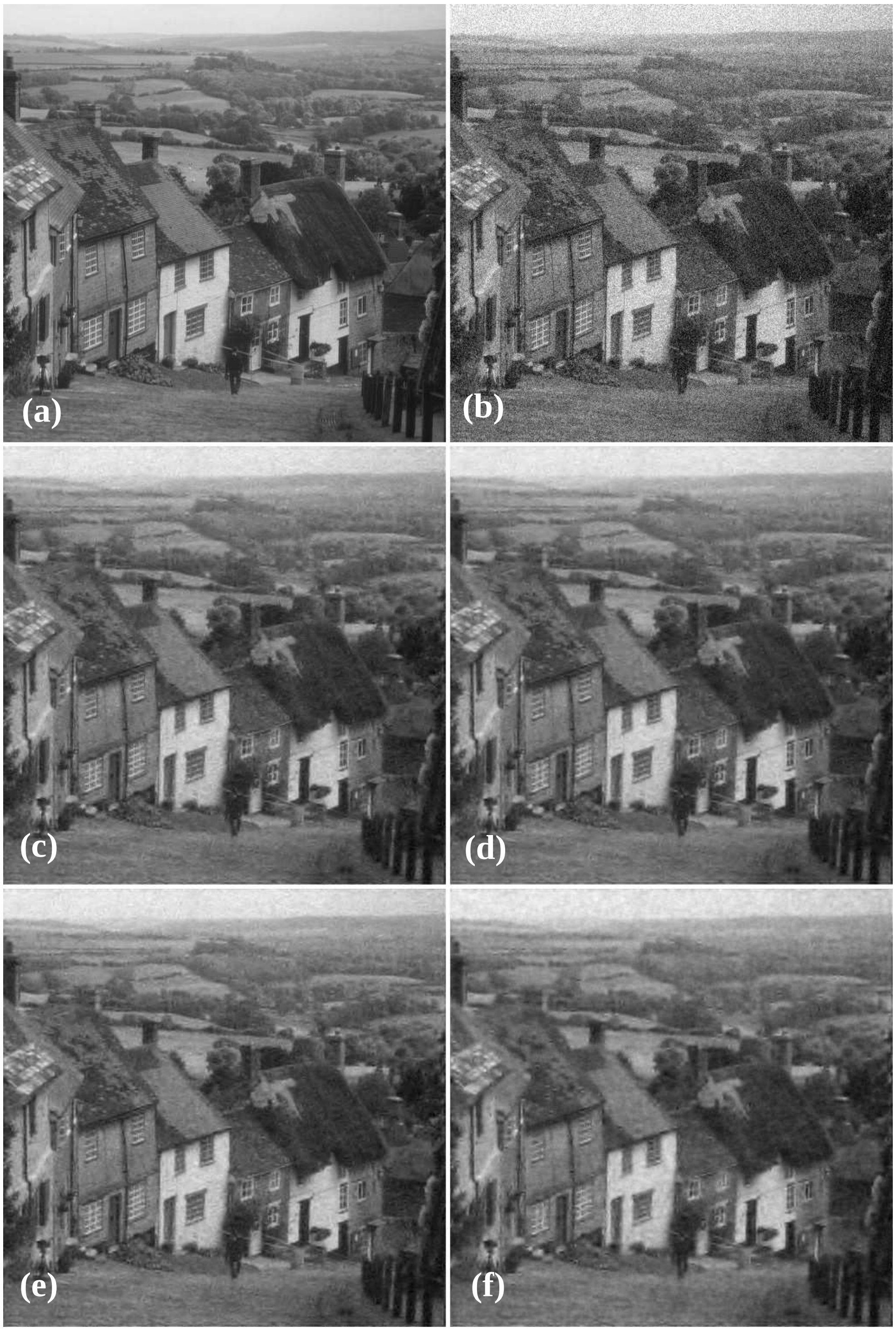


Fig. 10. Comparing the performance of the various methods for *goldhill* with $\sigma = 20$. (a) Original. (b) Noisy image, $\sigma = 20$. (c) *OracleShrink*. (d) *SureShrink*. (e) *BayesShrink*. (f) *BayesShrink* followed by MDLQ compression.

Fig. 11. Comparing the performance of the various methods for *lena* with $\sigma = 10$. (a) Original. (b) Noisy image, $\sigma = 10$. (c) *OracleShrink*. (d) *SureShrink*. (e) *BayesShrink*. (f) *BayesShrink* followed by MDLQ compression.

TABLE I

For various test images and σ values, lists MSE of (1) OracleShrink soft-thresholding, (2) SureShrink, (3) BayesShrink, (4) BayesShrink with MDLQ-compression, and (5) OracleThresh. The last column shows the bitrate (bits per pixel) of the compressed image of (4). Averaged over 5 runs.

	OracleShrink	*SureShrink*	*BayesShrink*	*BayesShrink* + Compress	*OracleThresh*	bitrate (bpp)
goldhill σ=10	41.28	42.26	**41.98**	59.21	54.11	1.055
σ=20	86.35	93.21	**88.59**	112.43	108.32	0.453
σ=30	124.80	128.98	**126.40**	155.60	152.22	0.270
σ=35	140.55	151.19	**141.97**	174.06	172.09	0.225
lena σ=10	28.31	**29.21**	29.65	40.71	34.93	0.747
σ=20	59.56	63.95	**61.73**	81.66	72.57	0.373
σ=30	89.74	94.13	**92.06**	119.37	110.16	0.234
σ=35	104.27	107.19	**106.45**	138.62	127.68	0.201
barbara σ=10	45.96	56.21	**51.27**	81.43	58.04	1.205
σ=20	118.11	**121.19**	121.52	170.20	150.94	0.852
σ=30	190.63	201.09	**192.60**	250.86	253.10	0.623
σ=35	226.30	246.07	**229.66**	289.77	300.11	0.516
baboon σ=10	66.37	85.76	**80.04**	129.71	90.27	1.343
σ=20	170.28	185.16	**180.15**	242.62	232.60	0.853
σ=30	263.09	**269.88**	270.18	332.66	349.62	0.583
σ=35	302.71	314.08	**309.92**	367.72	393.19	0.477

TABLE II

THE THRESHOLD VALUES OF *OracleShrink*, *SureShrink*, AND *BayesShrink*, RESPECTIVELY, (AVERAGED OVER 5 RUNS) FOR THE DIFFERENT SUBBANDS OF GOLDHILL, WITH NOISE STRENGTH $\sigma = 20$.

	Orientations		
Scales	HH	HL	LH
1 (finest)	84.90	37.36	38.04
2	37.12	21.16	18.50
3	17.31	11.51	8.45
4 (coarsest)	7.84	6.24	3.40

(a) *OracleShrink*: T_{OS}

	Orientations		
Scales	HH	HL	LH
1 (finest)	(95.66)	(95.66)	(95.66)
2	89.48	21.78	18.86
3	18.52	11.46	9.14
4 (coarsest)	8.44	5.58	3.56

(b) *SureShrink*: SURE or (universal)

	Orientations		
Scales	HH	HL	LH
1 (finest)	(95.85)	49.69	51.34
2	46.37	19.78	16.84
3	17.43	8.45	6.91
4 (coarsest)	7.03	2.86	3.13

(c) *BayesShrink*: T_B

TABLE III

THE VALUE OF m (AVERAGED OVER 5 RUNS) FOR THE DIFFERENT SUBBANDS OF GOLDHILL, WITH NOISE STRENGTH $\sigma = 20$.

	Orientations			
Scales	HH	HL	LH	LL
1 (finest)	0.0	4.0	4.2	
2	3.0	4.2	3.0	
3	3.4	5.2	5.0	
4 (coarsest)	5.0	15.4	9.8	27.4

TABLE IV

THE THRESHOLD VALUES OF *OracleShrink*, *SureShrink*, AND *BayesShrink*, RESPECTIVELY, (AVERAGED OVER 5 RUNS) FOR THE DIFFERENT SUBBANDS OF LENA, WITH NOISE STRENGTH $\sigma = 10$.

	Orientations		
Scales	HH	HL	LH
1 (finest)	25.73	14.13	19.16
2	11.21	7.66	10.18
3	6.15	3.58	5.89
4 (coarsest)	3.13	1.37	2.48

(a) *OracleShrink*: T_{OS}

	Orientations		
Scales	HH	HL	LH
1 (finest)	(48.99)	15.93	(48.99)
2	12.54	8.57	11.23
3	6.65	4.03	6.62
4 (coarsest)	3.30	1.06	2.79

(b) *SureShrink*: SURE or (universal)

	Orientations		
Scales	HH	HL	LH
1 (finest)	(48.83)	15.72	32.20
2	9.93	4.50	7.67
3	3.10	1.60	2.88
4 (coarsest)	1.21	0.58	1.25

(c) *BayesShrink*: T_B

TABLE V

THE VALUE OF m (AVERAGED OVER 5 RUNS) FOR THE DIFFERENT SUBBANDS OF LENA, WITH NOISE STRENGTH $\sigma = 10$.

	Orientations			
Scales	HH	HL	LH	LL
1 (finest)	0.0	5.2	5.6	
2	4.4	8.6	5.4	
3	9.0	11.6	8.0	
4 (coarsest)	18.6	31.8	14.2	45.0