Last modified: 30 March 1999.

Published as: Bayesian Inference in Wavelet Based Models
eds. P Müller and B Vidakovic
Chapter 18, pp 291--308
Lecture Notes in Statistics, volume 141
Springer-Verlag, New York, 1999.

Bayesian Denoising of Visual Images in the Wavelet Domain

Eero P. Simoncelli

The use of multi-scale decompositions has led to significant advances in representation, compression, restoration, analysis, and synthesis of signals. The fundamental reason for these advances is that the statistics of many natural signals, when decomposed in such bases, are substantially simplified. Choosing a basis that is adapted to statistical properties of the input signal is a classical problem. The traditional solution is principal components analysis (PCA), in which a linear decomposition is chosen to diagonalize the covariance structure of the input. The most well-known description of image statistics is that their Fourier spectra take the form of a power law [e.g., 1, 2, 3]. Coupled with a constraint of translation-invariance, this suggests that the Fourier transform is an appropriate PCA representation. Fourier and related representations are widely used in image processing applications. For example, the classical solution to the noise removal problem is the Wiener filter, which can be derived by assuming a signal model of decorrelated Gaussian-distributed coefficients in the Fourier domain.

Recently a number of authors have noted that statistics of order greater than two can be utilized in choosing a basis for images. Field [2, 4] noted that the coefficients of frequency subbands of natural scenes have much higher kurtosis than a Gaussian density. Recent work on so-called "independent components analysis" (ICA) has sought linear bases that optimize higher-order statistical measures [e.g., 5, 6]. Several authors have constructed optimal bases for images by optimizing such information-theoretic criterion [7, 8]. The resulting basis functions are oriented and have roughly octave bandwidth, similar to many of the most common multi-scale decompositions. A number of authors have explored the optimal choice of a basis from a library of functions based on entropy or other statistical criterion [e.g. 9, 10, 11, 12, 13].

In this chapter, we examine the empirical statistical properties of visual images within two fixed multi-scale bases, and describe two statistical models for the coefficients in these bases. The first is a non-Gaussian marginal model, previously described in [14]. The second is a joint non-Gaussian

Research supported by NSF CAREER grant MIP-9796040.

Markov model for wavelet subbands, previous versions of which have been described in [15, 16]. We demonstrate the use of each of these models in Bayesian estimation of an image contaminated by additive Gaussian white noise.

1 Marginal Statistical Model

A number of authors have observed that wavelet subband coefficients have highly non-Gaussian statistics [e.g., 17, 4, 14]. The intuitive explanation for this is that images typically have spatial structure consisting of smooth areas interspersed with occasional edges or other abrupt transitions. The smooth regions lead to near-zero coefficients, and the structures give occasional large-amplitude coefficients.

As an example, histograms for subbands of separable wavelet decompositions of several images are plotted in figure 1[1]. These densities may be accurately modeled with a two-parameter density function of the form [17, 14]:

$$\mathcal{P}_c(c; s, p) = \frac{e^{-|c/s|^p}}{Z(s,p)}, \tag{1.1}$$

where the normalization constant is $Z(s,p) = 2\frac{s}{p}\Gamma(\frac{1}{p})$. Each graph in figure 1 includes a dashed curve corresponding to the best fitting instance of this density function, with the parameters $\{s, p\}$ estimated by maximizing the likelihood of the data under the model. For subbands of images in our collection, values of the exponent p typically lie in the range $[0.5, 0.8]$. The density model fits the histograms remarkably well, as indicated by the relative entropy measures given below each plot.

1.1 Bayesian denoising: Marginal model

Consider an image whose pixels are contaminated with i.i.d. samples of additive Gaussian noise. Because the wavelet transform is orthonormal, the noise is also Gaussian and white in the wavelet domain. Thus, each coefficient in the wavelet expansion of the noisy image is written as $y = c + n$, where c is drawn from the marginal density given in equation (1.1), and n is Gaussian.

A standard estimator for c given the corrupted observation y is the maximum a posteriori (MAP) estimator:

$$\hat{c}(y) \quad = \quad \arg\max_c \mathcal{P}_{c|y}(c|y) \tag{1.2}$$

[1] The specific wavelet decomposition used for these examples is described in section 2.2.

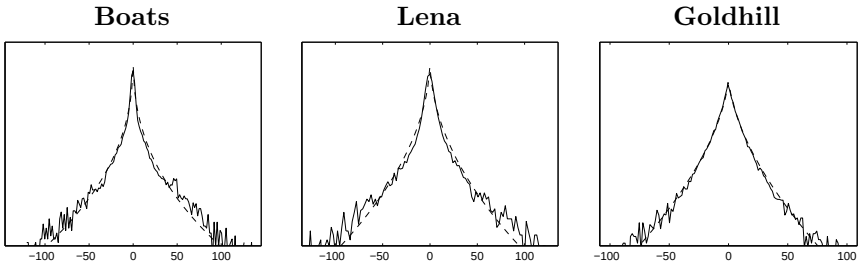


FIGURE 1. Examples of 256-bin coefficient histograms for a single vertical wavelet subband of three images, plotted in the log domain. All images are size $512x512$. Also shown (dashed lines) are fitted model densities corresponding to equation (1.1). Below each histogram is the maximum-likelihood value of p used for the fitted model density, and the relative entropy (Kullback-Leibler divergence) of the model and histogram, as a fraction of the total entropy of the histogram.

$$= \arg\max_c \mathcal{P}_{y|c}(y|c)\mathcal{P}_c(c) \tag{1.3}$$

$$= \arg\max_c \mathcal{P}_n(y-c)\mathcal{P}_c(c) \tag{1.4}$$

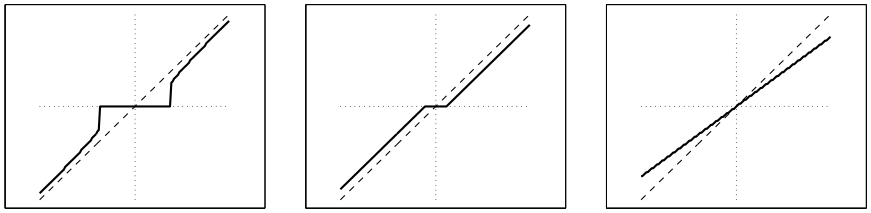
where Bayes' rule allows us to write this in terms of the probability densities of the noise ($\mathcal{P}_n$) and the prior density of the signal coefficient ($\mathcal{P}_c$). In order to use this equation to estimate the original signal value c, we must know both density functions.

Figure 2 shows a set of (numerically computed) MAP estimators for the model in equation (1.1) with different values of the exponent p, assuming a Gaussian noise density. In the special case of $p = 2$ (i.e., Gaussian source density), the estimator assumes the well-known linear form:

$$\hat{c}(y) = \frac{\sigma_c^2 \; y}{\sigma_c^2 + \sigma_n^2}, \tag{1.5}$$

estimators for other values of p are nonlinear: the $p = 0.5$ function resembles a hard thresholding operator, and $p = 1$ resembles a soft thresholding operator. Donoho has shown that these types of shrinkage operator are nearly minimax optimal for some classes of regular function (e.g., Besov) [18]. Other authors have established connections of these these results with statistical models [19, 20]. In addition, thresholding techniques are widely used in the television and video engineering community, where they are known as "coring" [e.g., 21, 22, 23]. For example, most consumer VCR's use a simple coring technique to remove magnetic tape noise.

If one wishes to minimize squared error, the mean of the posterior distribution provides an optimal estimate of the coefficient c, given a measure-



$p = 0.5$ $p = 1.0$ $p = 2.0$

FIGURE 2. MAP estimators for the model given in equation (1.1), with three different exponents. The noise is additive and Gaussian, with variance one third that of the signal. Dashed line indicates the identity function.

ment of y:

$$\begin{aligned} \hat{c}(y) &= \int dc\, \mathcal{P}_{c|y}(c|y)\, c \\ &= \frac{\int dc\, \mathcal{P}_{y|c}(y|c)\, \mathcal{P}_c(c)\, c}{\int dc\, \mathcal{P}_{y|c}(y|c)\, \mathcal{P}_c(c)} \\ &= \frac{\int dc\, \mathcal{P}_n(y-c)\, \mathcal{P}_c(c)\, c}{\int dc\, \mathcal{P}_n(y-c)\, \mathcal{P}_c(c)}. \end{aligned} \tag{1.6}$$

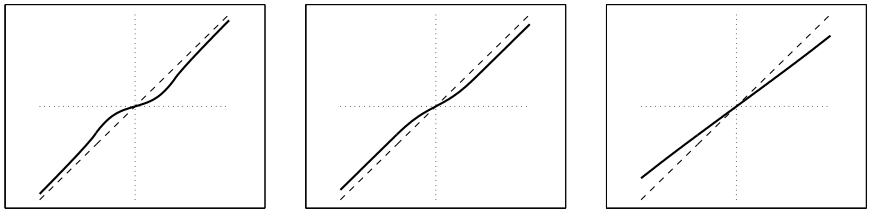
The denominator is the pdf of the noisy observation y, computed by marginalizing the convolution of the noise and signal pdf's.

Figure 3 shows (numerically computed) Bayesian least-squares estimators for the model of equation (1.1), with three different values of the exponent p. Again, for the special case of $p = 2$ the estimator is linear and of the form of equation (1.5). As with the MAP estimators, smaller values of p produce a nonlinear shrinkage operator, somewhat smoothed in comparison to those of figure 2. In particular, for $p = 0.5$ (which is well-matched to wavelet marginals such as those shown in figure 1), the estimator preserves large amplitude values and suppresses small amplitude values. This is intuitively sensible: given the substantial prior probability mass at $c = 0$, small values of y are assumed to have arisen from a value of $c = 0$.

The quality of a denoising algorithm will depend on the exponent p. To quantify this, figure 4 shows the (numerically computed) error variance for the Bayesian least-squares estimate (see figure 3), as a function of p. Note that the error variance drops significantly for values of p less than one.

In practice, one must estimate the parameters $\{s, p\}$ and σ_n from the noisy collection of coefficients $\{y_k\}$. A simple solution is a maximum likelihood estimator:

$$\{\hat{s}, \hat{p}, \hat{\sigma}_n\} = \arg \max_{\{s,p,\sigma_n\}} \prod_k \mathcal{P}_y(y_k; s, p, \sigma_n)$$



$p = 0.5$ $p = 1.0$ $p = 2.0$

FIGURE 3. Bayesian least-squares estimators for the model given in equation (1.1), with three different exponents, p. The noise is additive and Gaussian, with variance one third that of the signal. Dashed line indicates the identity function.

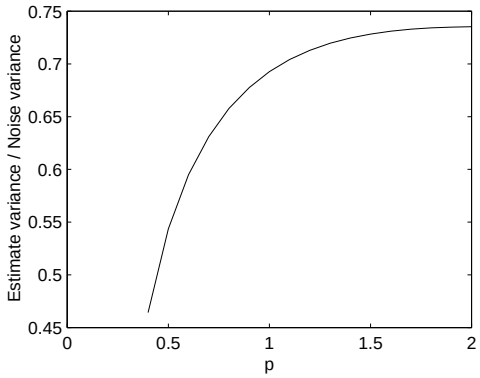


FIGURE 4. Error variance of the Bayes least-squares estimator relative to the noise variance, as a function of the model parameter p of equation (1.1). Noise variance σ_n was held constant at one third of the signal variance.

$$= \arg \max_{\{s,p,\sigma_n\}} \prod_k \int dc \, e^{-|c/s|^p} \, e^{-(y_k - c)^2/2\sigma_n^2} \qquad (1.7)$$

where the product is taken over all coefficients within the subband. In practice, both the integration and the optimization are performed numerically. Furthermore, in the examples shown in section 3, we assume σ_n is known, and optimize only over $\{s, p\}$. As a starting point for the optimization, we solve for the parameter pair $\{s, p\}$ corresponding to a density with kurtosis and variance matching those of the histogram [as in 14].

2 Joint Statistical Model

In the model of the previous section, we treated the wavelet coefficients as if they were independent. Empirically, orthonormal wavelet coefficients

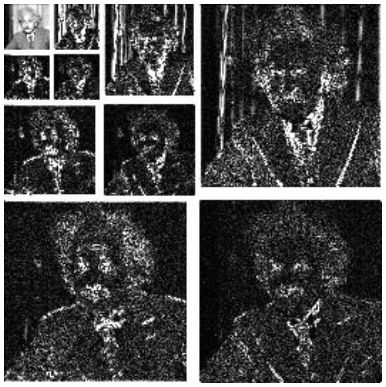


FIGURE 5. Coefficient magnitudes of a wavelet decomposition. Shown are absolute values of subband coefficients at three scales, and three orientations of a separable wavelet decomposition of the "Einstein" image.

are found to be fairly well decorrelated. Nevertheless, it is quite evident that wavelet coefficients of images are *not* statistically independent. Figure 5 shows the magnitudes (absolute values) of coefficients in a four-level separable wavelet decomposition. In particular, previous work has shown that large-magnitude coefficients tend to occur near each other within subbands, and also occur at the same relative spatial locations in subbands at adjacent scales, and orientations [e.g., 15, 16].

As an example, consider two coefficients representing horizontal information at adjacent scales, but the same spatial location of the "Boats" image. Figure 6A shows the conditional histogram $\mathcal{H}(c|p)$ of the "child" coefficient conditioned on a coarser-scale "parent" coefficient. The histogram illustrates several important aspects of the relationship between the two coefficients. First, they are (second-order) decorrelated, since the expected value of c is approximately zero for all values of p. Second, the variance of the conditional histogram of c clearly depends the value of p. Thus, although c and p are uncorrelated, *they are still statistically dependent*. Furthermore, this dependency cannot be eliminated through further linear transformation.

The structure of the relationship between c and p becomes more apparent upon transforming to the log domain. Figure 6B shows the conditional histogram $\mathcal{H}\left(\log_2(c^2)|\log_2(p^2)\right)$ The right side of the distribution is unimodal and concentrated along a unit-slope line. This suggests that in this region, the conditional expectation, $\mathbb{E}(c^2|p^2)$, is approximately proportional to p^2. Furthermore, vertical cross sections (i.e., conditional histogram for a fixed value of p^2) have approximately the same shape for different values of p^2. Finally, the left side of the distribution is concentrated about a horizontal line, suggesting that c^2 is independent of p^2 in this region.

The form of the histograms shown in figure 6 is surprisingly robust across

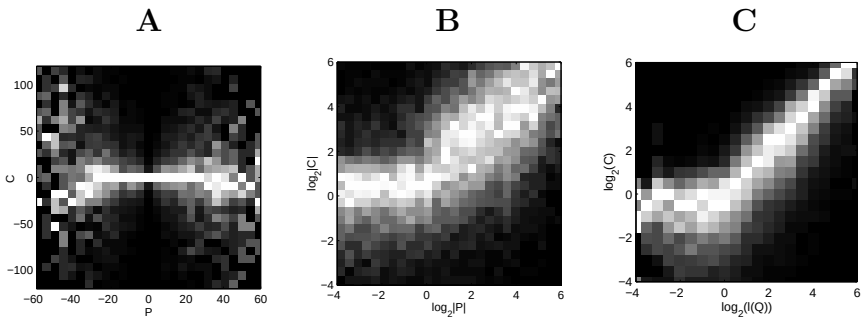


FIGURE 6. Conditional histograms for a fine scale horizontal coefficient. Brightness corresponds to probability, except that each column has been independently rescaled to fill the full range of display intensities. **A:** Conditioned on the parent (same location and orientation, coarser scale) coefficient. Data are for the "Boats" image. **B:** Same as **A**, but in the log domain. **C** Conditioned on a linear combination of neighboring coefficient magnitudes.

a wide range of images. Furthermore, the qualitative form of these statistical relationships also holds for pairs of coefficients at adjacent spatial locations ("siblings"), adjacent orientations ("cousins"), and adjacent orientations at a coarser scale ("aunts"). Given the linear relationship between the squares of large-amplitude coefficients and the difficulty of characterizing the full density of a coefficient conditioned on its neighbors, we've examined a linear predictor for the squared coefficient. Figure 6C shows a histogram of $\log_2(c^2)$ conditioned on a linear combination of the squares of eight adjacent coefficients in the same subband, two coefficients at other orientations, and a coefficient at a coarser scale. The linear combination is chosen to be least-squares optimal (see equation (1.9)). The histogram is similar to the single-band conditional histogram of figure 6B, but the linear region is extended and the conditional variance is significantly reduced.

The form of these observations suggests a simple Markov model, in which the density of a coefficient, c, is conditionally Gaussian with variance a linear function of the squared coefficients in a local neighborhood:

$$\mathcal{P}\left(c \mid \vec{p}\right) \;=\; \mathcal{N}\left(0; \sum_k w_k p_k^2 + \alpha^2\right). \tag{1.8}$$

Here, the neighborhood $\{p_k\}$ consists of coefficients at other orientations and adjacent scales, as well as adjacent spatial locations. Note that although we utilize a normal distribution, this is not a jointly Gaussian density in the traditional sense, since the *variance* rather than the mean is dependent on the neighborhood. Figure 7 shows a set of conditional histograms, and the best-fitting instantiation of the model in equation (1.8). The fits are seen to be reasonably good, as indicated by the low relative entropy values.

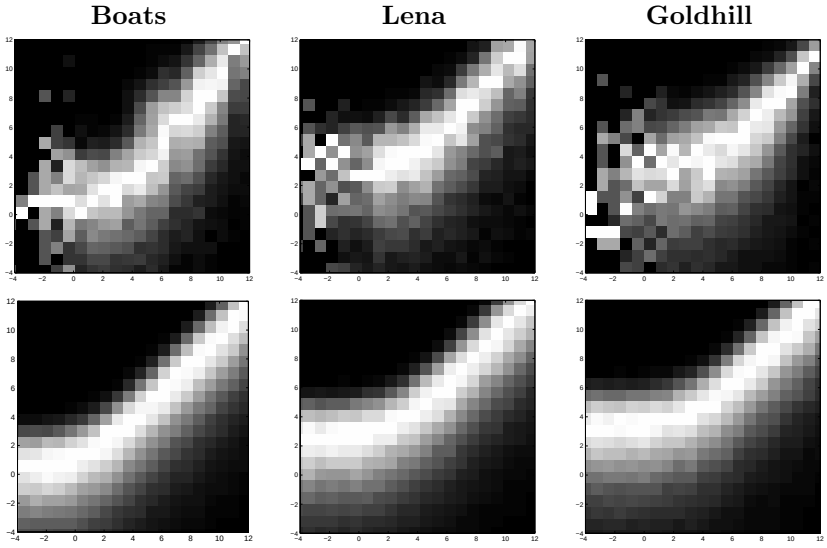


FIGURE 7. Top: Examples of log-domain conditional histograms for the second-level horizontal subband of different images, conditioned on an optimal linear combination of coefficient magnitudes from adjacent spatial positions, orientations, and scales. **Bottom**: Model of equation (1.8) fitted to the conditional histograms in the left column. Intensity corresponds to probability, except that each column has been independently rescaled to fill the full range of intensities.

We have used variants of this model for applications of compression [15] and texture synthesis [24].

2.1 Bayesian denoising: Joint model

As in the previous section, assume a coefficient is contaminated with Gaussian white noise: $y = c+n$. If we assume the neighbor coefficients are known, the conditionally Gaussian form of equation (1.8) leads to a linear Bayesian estimator:

$$\hat{c}(y) = \frac{\sum_k w_k p_k^2 + \alpha^2}{\sum_k w_k p_k^2 + \alpha^2 + \sigma_n^2} y.$$

In a more realistic implementation, we must estimate c given the noisy observations of the neighbors. A complete solution for this problem is difficult, since the conditional density of the variance of the clean coefficient given the noisy neighbors cannot be computed in closed form. A numerical solution should be feasible, but for the purposes of the current paper, we instead choose to utilize the marginal model estimator from the previous section.

Specifically, we first compute estimates of the coefficients in a subband,

$\hat{c}$, using equation (1.6). We then use these estimated coefficients to estimate the weight parameters, $\{w_k\}$, and the constant, α, by minimizing the squared error:

$$\{\hat{w}, \hat{\alpha}\} = \arg \min_{\{\vec{w},\alpha\}} \mathbb{E}\left[\hat{c}^2 - \sum_k w_k \hat{p}_k^2 - \alpha^2\right]^2. \tag{1.9}$$

Note that we are using the marginal estimates of both the coefficient *and* the neighbors. We have also implemented (numerically) a maximum likelihood solution, but it was found to be computationally expensive and did not yield any improvement in performance.

Given $\hat{w}$, the joint model estimate, $\hat{\hat{c}}(y)$, is computed from the noisy observation, y, using the marginal estimates of the neighbors:

$$\hat{\hat{c}}(y) = \frac{\sum_k \hat{w}_k \hat{p}_k^2 + \hat{\alpha}^2}{\sum_k \hat{w}_k \hat{p}_k^2 + \hat{\alpha}^2 + \sigma_n^2}\, y. \tag{1.10}$$

Although clearly sub-optimal, this estimate is easily computed and gives reasonable results.

2.2 Choice of Basis

As mentioned in the introduction, a number of recent researchers have derived wavelet-like bases for images using information-theoretic optimality criterion. Here, we compare the denoising abilities of two different types of discrete multi-scale basis.

The first is a separable critically-sampled 9-tap quadrature mirror filter (QMF) decomposition, based on filters designed in [25]. This is a linear-phase (symmetric) approximation to an orthonormal wavelet decomposition. The lowpass filter samples are:

$$\begin{aligned} l[n] \;=\; & [0.028074, -0.060945, -0.073387, 0.41473, 0.79739, \\ & 0.41473, -0.073387, -0.060945, 0.028074]. \end{aligned}$$

The highpass filter is obtained via $h[n] = (-1)^n l[N-n+1]$, and the system diagram is shown in figure 8. Compared with orthonormal wavelets, this decomposition has the advantage that the basis functions are symmetric. The drawback is that the system does not give perfect reconstruction: the filters are designed to optimize a residual function. This is not a serious problem for applications such as the one discussed in this chapter, since reconstruction signal-to-noise ratios (SNRs) are typically about 55dB.

The second decomposition is known as a *steerable pyramid* [26]. In this decomposition, the image is subdivided into subbands using filters that are polar-separable in the Fourier domain. In scale, the subbands have octave bandwidth with a functional form constrained by a recursive system diagram. In orientation, the functional form is chosen so that the set filters

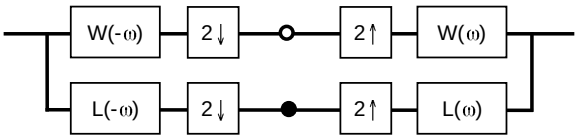


FIGURE 8. Single-scale system diagram for a critically-sampled QMF or wavelet decomposition, in one dimension. Boxes correspond to convolution, downsampling, and upsampling operations. Two-dimensional decomposition is achieved by applying the one-dimensional decomposition in the vertical direction, and then to both resulting subbands in the horizontal direction. Multi-scale decompositions are constructed by inserting the system into itself at the location of the filled circle.

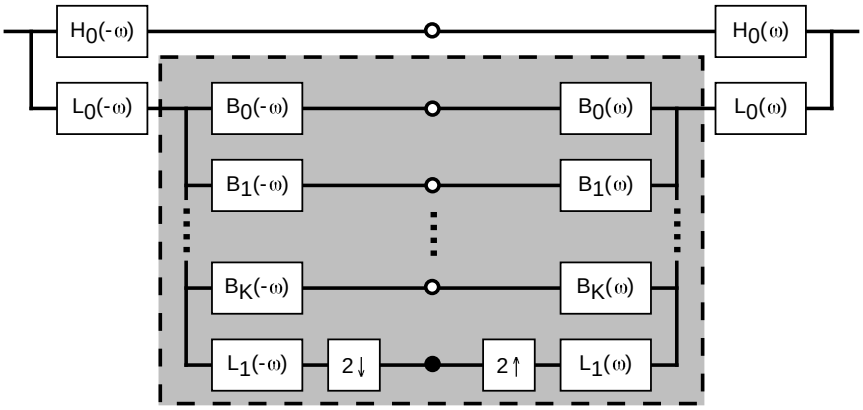


FIGURE 9. Single-scale system diagram for a steerable pyramid. Multi-scale decompositions are constructed by inserting the portion of the system within the gray region at the location of the filled circle.

at a given scale span a rotation-invariant subspace. The decomposition can performed with any number of orientation bands, K, each of orientation bandwidth $2\pi/K$ radians. The full two-dimensional transform is overcomplete by a factor of $4K/3$, and is a tight frame (i.e., the matrix corresponding to the inverse transformation is equal to the transpose of the forward transformation matrix). Spatial subsampling of each subband respects the Nyquist criterion, and thus the representation is translation-invariant (free of aliasing). An idealized system diagram is shown in figure 9.

The transform is implemented using a set of oriented filters that are polar-separable when expressed in the Fourier domain:

$$F_{n,k}(r,\theta) = B_n(r)G_k(\theta), \quad n \in [0,M], k \in [0,K-1],$$

where

$$B_n(r) = \cos\left(\frac{\pi}{2}\log_2\left(\frac{2^n r}{\pi}\right)\right)$$

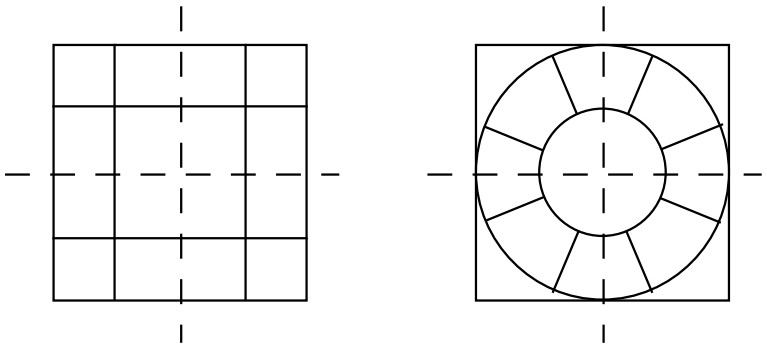


FIGURE 10. Idealized partition of frequency domain associated with each decomposition. Each axis covers the range $[-\pi, \pi]$ radians/pixel. Left: Separable QMF or wavelet. Right: Steerable pyramid, $K = 4$.

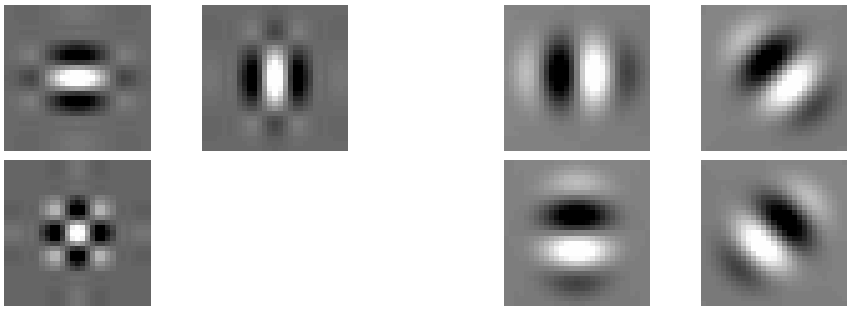


FIGURE 11. Basis functions at a single scale. Left: QMF decomposition. Right: 4-orientation steerable pyramid.

$$G_k(\theta) = \begin{cases} \left[\cos(\theta - \frac{\pi k}{K})\right]^{K-1}, & |\theta - \frac{\pi k}{K}| < \frac{\pi}{2} \\ 0, & \text{otherwise,} \end{cases}$$

where r, θ are polar frequency coordinates. Subbands are subsampled by a factor of 2^n along both axes. In addition, one must retain the (non-oriented) lowpass residual band, which is computed using the following filter:

$$L(r) \quad = \quad \begin{cases} \cos\left(\frac{\pi}{2}\log_2(\frac{2^{(M+1)}r}{\pi})\right), & r \in [\frac{\pi}{2^{M+1}}, \frac{\pi}{2^M}] \\ 1 & r < \frac{\pi}{2^{M+1}} \\ 0 & r > \frac{\pi}{2^M}. \end{cases}$$

Figure 10 shows the idealized frequency partition of the two decompositions, figure 11 shows the basis functions at a single scale, and figure 12 shows the decomposition of the Einstein image using the two bases.

Figure 13 shows a scatter plot of estimated values of p for the images shown in figure 1. Note that the values for some of the separable QMF bands

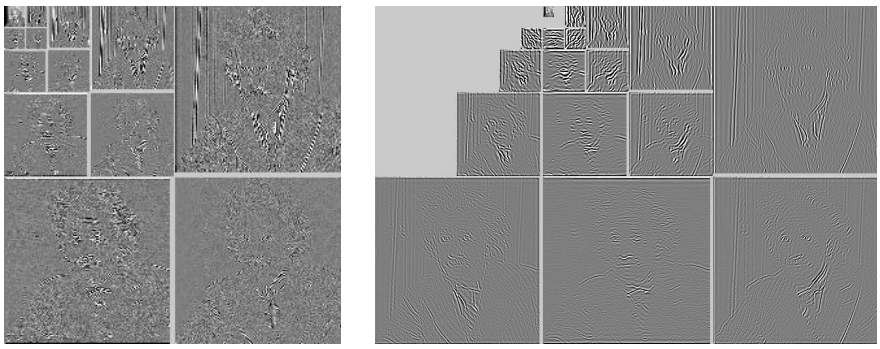


FIGURE 12. Example multi-scale decompositions. Left: separable QMF decomposition, a linear-phase approximation to an orthonormal wavelet decomposition. Right: steerable pyramid, with $K = 4$ orientation bands.

are quite high (particularly, band 3, which contains the mixed diagonals). On average, the steerable pyramid values less than those of the separable critically-sampled system. This is consistent with the preference for oriented basis functions, as mentioned in the introduction.

In addition to the smaller values of p, another advantage of the steerable pyramid in the context of denoising is the translation-invariance property. Previous work has emphasized the importance of translation-invariance for image processing tasks such as denoising [26, 11, 12, 14]. One drawback of this representation is that our assumption of orthonormality is violated. In particular, the representation is heavily overcomplete and thus there are strong linear dependencies between the coefficients. The marginal model of section 1 assumes that the coefficients are statistically independent, and the joint model of section 2 assumes that they are decorrelated.

3 Results

In this section, we show examples of image denoising using the two models described in previous sections. In all cases, we assume the noise variance, σ_n^2, is known. Gaussian noise of this variance, truncated to a range of three standard deviations, is added to the original image. This contaminated image is transformed to the relevant basis, the appropriate estimator is applied to all coefficients within each subband, and then the transformation is inverted. The estimators are computed as follows:

- Linear estimator:
  1. Estimate $\sigma_c^2 \approx \max\{0, \mathbb{E}\left(c^2 - \sigma_n^2\right)\}$.
  2. $\hat{c}(y) = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_n^2} y$.

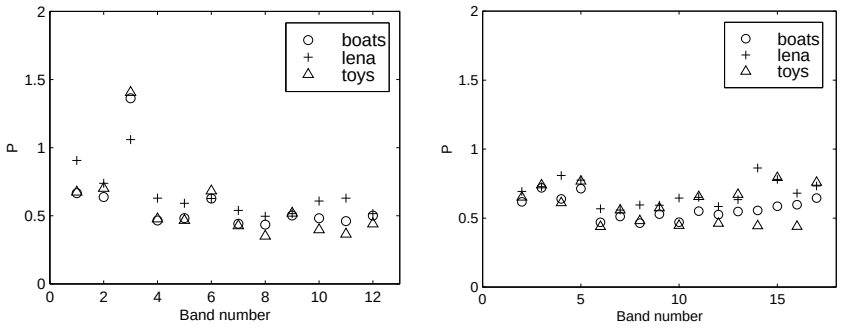


FIGURE 13. Values of p for subbands from the 512×512 images analyzed in figure 1. Left: Subbands of separable QMF pyramid. Right: Subbands of steerable pyramid ($K = 4$ orientations). Subband numbering runs from high to low frequency. Orientation bands of the separable decomposition are ordered (vertical,horizontal,mixed-diagonal), and orientation bands of the steerable pyramid start with vertical and proceed counterclockwise.

- Threshold estimator:

  1. $t = 3\sigma_n$
  2. $\hat{c}(y) = \begin{cases} y, & |y| > t \\ 0, & \text{otherwise} \end{cases}$

- Bayesian marginal (coring) estimator:

  1. Compute parameter estimates $\{\hat{s}, \hat{p}\}$ by maximizing likelihood of the subband data (equation (1.7)).
  2. Compute the conditional mean estimator $f(y)$ numerically using equation (1.6).
  3. $\hat{c}(y) = f(y)$

- Bayesian joint estimator:

  1. Compute $\hat{c}(y)$ for all subbands using the Bayesian marginal estimator.
  2. Estimate weights $\hat{w}$ and $\hat{\alpha}$ using equation (1.9).
  3. $\hat{\hat{c}}(y) = \frac{\sum_k \hat{w}_k \hat{p}_k^2 + \hat{\alpha}^2}{\sum_k \hat{w}_k \hat{p}_k^2 + \hat{\alpha}^2 + \sigma_n^2}\ y.$

All expectations are estimated by summing spatially. For the joint estimator, we use a neighborhood consisting of the 12 nearest spatial neighbors (within the same subband), the 5 nearest cousin coefficients (in other orientation bands at the same scale), the 9 nearest parent coefficients (in the

decomposition type	noisy image	Estimator			
		Linear	Threshold	BayesCore	BayesJoint
QMF-9	1.59	8.26	8.09	9.66	10.58
	4.80	9.71	10.10	11.64	12.31
	8.99	12.02	12.46	14.21	14.49
	13.03	14.81	14.65	16.61	16.62
Spyr (4 ori)	1.59	9.28	10.12	10.35	10.96
	4.80	10.60	11.71	11.98	12.61
	9.02	12.58	14.04	14.19	14.81
	13.06	14.96	15.89	16.50	16.99
Spyr (6 ori)	1.59	9.34	10.37	10.55	11.12
	4.80	10.67	12.03	12.18	12.74
	9.02	12.66	14.23	14.43	14.90
	13.06	15.02	16.28	16.71	17.09

TABLE 1.1. Denoising results for four estimators, three different decompositions, and four different levels of additive Gaussian noise added to the "Einstein" image. All values indicate signal-to-noise ratio in decibels ($10 \log_{10}$(signal variance/error variance)).

adjacent subband of coarser scale), a single aunt (from each orientation at the adjacent coarser scale), and a single grandparent.

Table 3 shows signal-to-noise ratios (SNRs) for all four algorithms, applied to all three decompositions, at four different contamination levels. Note that Bayesian algorithms outperform the other two techniques for all examples. Also note that the steerable pyramid decompositions significantly outperform the separable QMF decomposition, and the six-orientation decomposition shows a noticeable improvement over the four-orientation decomposition.

Finally, figures 14 and 15 show some example images. Figure 14 shows results using the separable decomposition. The Bayesian results appear both sharper (because high-amplitude coefficients are preserved) and less noisy (because low-amplitude coefficients are suppressed) than the linear estimator. The aliasing artifacts of the critically sampled transform are most evident with the thresholding estimator, and least evident in the Bayes joint estimator.

Figure 15 shows results using the 4-orientation steerable pyramid. Note that although the use of this decomposition eliminates the aliasing artifacts and produces higher SNR results, the results now look more blurred. The results can be made more visually appealing by a subsequent sharpening operation, although this reduces the SNR.

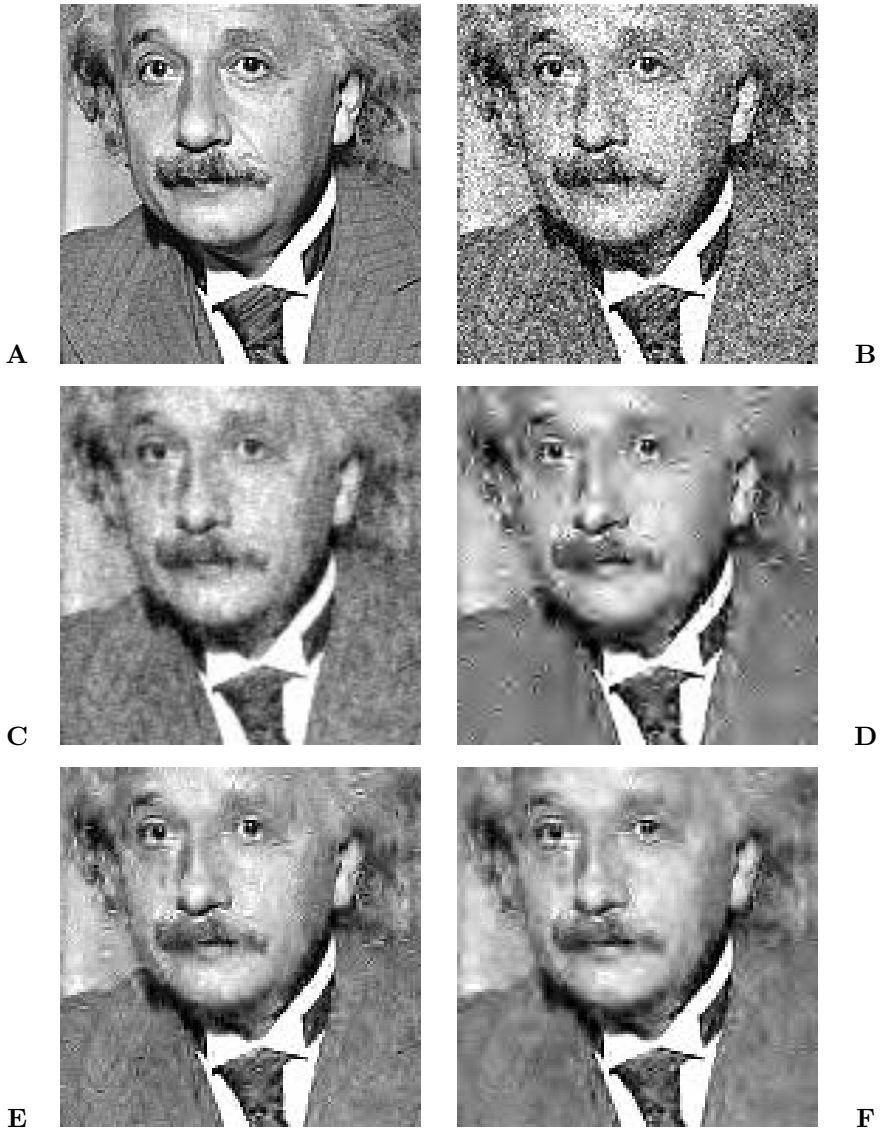


FIGURE 14. Cropped denoising results using an (approximately) orthonormal separable decomposition. **A**: Original "Einstein" image (cropped). **B**: Noisy image (SNR = 4.8dB). **C**: Linear least-squares estimator. **D**: Optimal thresholding. **E**: Bayes - marginal model. **F**: Bayes - joint model.

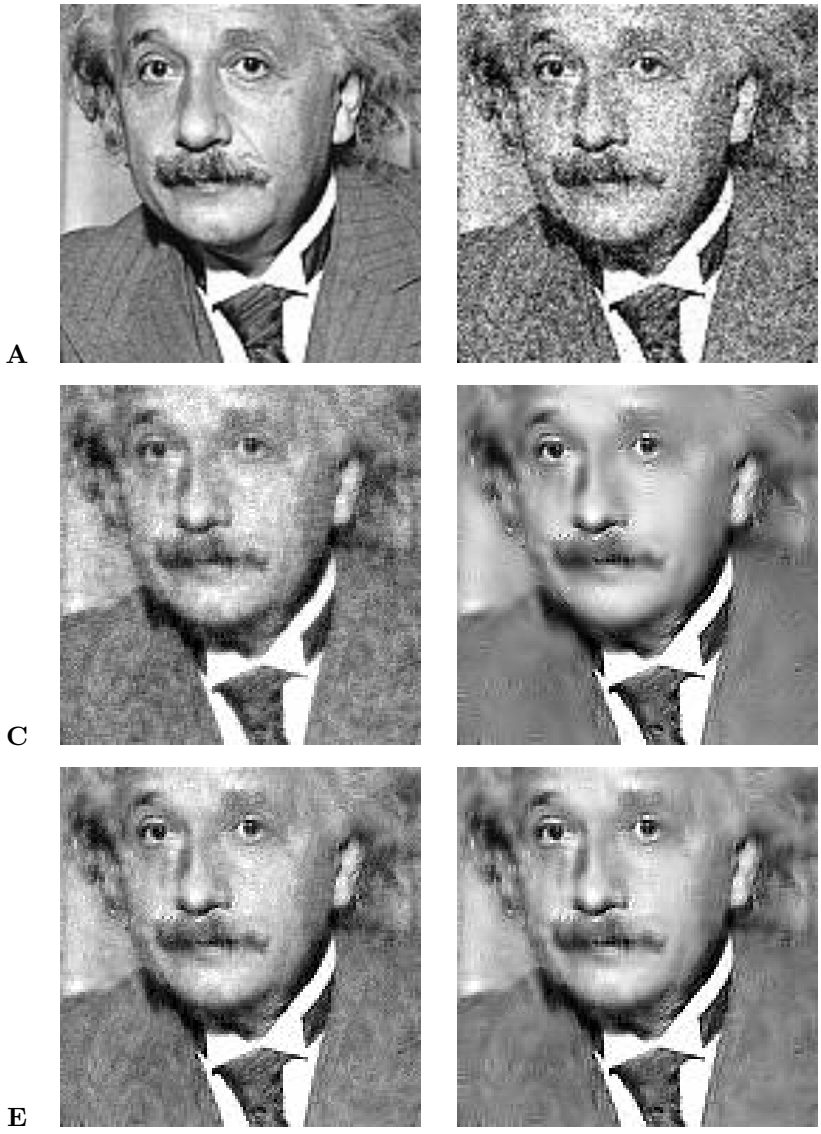


FIGURE 15. Cropped denoising results using a 4-orientation steerable pyramid decomposition. **A**: Original "Einstein" image (cropped). **B**: Noisy image (SNR = 4.8dB). **C**: Linear least-squares estimator. **D**: Optimal thresholding. **E**: Bayes - marginal model. **F**: Bayes - joint model.

4 Conclusion

We have described two non-Gaussian density models for visual images, and used them to develop nonlinear Bayesian estimators that outperform classical linear estimators and simple thresholding estimators.

We have implemented these estimators in the context of two different multi-scale decompositions. The results obtained with the overcomplete steerable pyramid bases are superior to the separable QMF basis, due to translation-invariance and more kurtotic statistics. This is true in spite of the fact that the statistical model is no longer correct: the use of an overcomplete basis creates linear dependencies between coefficients. It would be interesting to compare these results to other translation-invariant denoising schemes, such as the cycle-spinning approach of [27]. It would also be interesting to explore whether the results can be improved by adaptively choosing an optimal basis. Such optimization could be done over a collection of images drawn from a particular class, or individually for each image (assuming spatial stationarity).

A number of improvements should be made to the statistical models, and the Bayesian estimators described in this paper. In particular, the marginal densities that come from integrating the joint model density are inconsistent with those of the generalized Gaussian marginal model of section 1. Although the empirical evidence for the linear dependency of coefficient variance on a single neighbor is quite strong, the linear estimate based on the full neighborhood needs to be validated. A full joint Bayesian estimator, which estimates both the density parameters and the clean coefficient based on the noisy observations, should be implemented numerically. The distortion model should be extended to include spatial blurring. A fully blind denoising technique (i.e., including estimation of σ_n) should also be explored.

Finally, we note the need for a measure of image distortion that adequately reflects human perceptual salience. Although the Bayesian denoising results in figure 15 are excellent according to a squared error measure, informal questioning suggests that most observers prefer a sharper image, even if it contains more noticeable artifacts.

5 References

[1] Alex Pentland. Fractal based description of natural scenes. *IEEE Trans. PAMI*, 6(6):661–674, 1984.

[2] D J Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4(12):2379–2394, 1987.

[3] D L Ruderman and W Bialek. Statistics of natural images: Scaling in the woods. *Phys. Rev. Letters*, 73(6), 1994.

[4] D J Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.

[5] J F Cardoso. Source separation using higer order moments. In *ICASSP*, pages 2109–2112, 1989.

[6] P Common. Independent component analysis, a new concept? *Signal Process.*, 36:387–314, 1994.

[7] B A Olshausen and D J Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7:333–339, 1996.

[8] A J Bell and T J Sejnowski. Learning the higher-order structure of a natural sound. *Network: Computation in Neural Systems*, 7:261–266, 1996.

[9] R R Coifman and M V Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. Info. Theory*, IT-38:713–718, March 1992.

[10] Stephane Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Proc.*, December 1993.

[11] D L Donoho and I M Johnstone. Ideal denoising in an orthogonal basis chosen from a library of bases. *C.R. Acad. Sci.*, 319:1317–1322, 1994.

[12] M V Wickerhauser. *Adapted Wavelet Analysis: From Theory to Software*. A K Peters, Wellesley, MA, 1994.

[13] J C Pesquet, H Krim, D Leporini, and E Hamman. Bayesian approach to best basis selection. In *Proc Int'l Conf Acoustics, Speech and Signal Proc*, pages 2634–2638, Atlanta, May 1996.

[14] E P Simoncelli and E H Adelson. Noise removal via Bayesian wavelet coring. In *Third Int'l Conf on Image Proc*, volume I, pages 379–382, Lausanne, September 1996. IEEE Sig Proc Society.

[15] R W Buccigrossi and E P Simoncelli. Image compression via joint statistical characterization in the wavelet domain. Technical Report 414, GRASP Laboratory, University of Pennsylvania, May 1997. Accepted (3/99) for publication in IEEE Trans Image Processing.

[16] E P Simoncelli. Statistical models for images: Compression, restoration and synthesis. In *31st Asilomar Conf on Signals, Systems and Computers*, pages 673–678, Pacific Grove, CA, November 1997. IEEE Computer Society. Available at: ftp://ftp.cns.nyu.edu/pub/eero/simoncelli97b.ps.gz.

[17] S G Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Pat. Anal. Mach. Intell.*, 11:674–693, July 1989.

[18] D Donoho and I Johnstone. Adapting to unknown smootness via wavelet shrinkage. *J American Stat Assoc*, 90(432), December 1995.

[19] F Abramovich, T Sapatinas, and B W Silverman. Wavelet thresholding via a bayesian approach. *J R Stat Soc B*, 60:725–749, 1998.

[20] D Leporini and J C Pesquet. Multiscale regularization in besov spaces. In *31st Asilomar Conf on Signals, Systems and Computers*, Pacific Grove, CA, November 1998.

[21] J P Rossi. *JSMPTE*, 87:134–140, 1978.

[22] B. E. Bayer and P. G. Powell. A method for the digital enhancement of unsharp, grainy photographic images. *Adv in Computer Vision and Im Proc*, 2:31–88, 1986.

[23] J. M. Ogden and E. H. Adelson. Computer simulations of oriented multiple spatial frequency band coring. Technical Report PRRL-85-TR-012, RCA David Sarnoff Research Center, April 1985.

[24] E Simoncelli and J Portilla. Texture characterization via joint statistics of wavelet coefficient magnitudes. In *Fifth IEEE Int'l Conf on Image Proc*, volume I, Chicago, October 4-7 1998. IEEE Computer Society.

[25] E P Simoncelli and E H Adelson. Subband transforms. In John W Woods, editor, *Subband Image Coding*, chapter 4, pages 143–192. Kluwer Academic Publishers, Norwell, MA, 1990.

[26] E P Simoncelli, W T Freeman, E H Adelson, and D J Heeger. Shiftable multi-scale transforms. *IEEE Trans Information Theory*, 38(2):587–607, March 1992. Special Issue on Wavelets.

[27] R R Coifman and D L Donoho. Translation-invariant de-noising. Technical Report 475, Statistics Department, Stanford University, May 1995.

Errata, 14 May 1999

- Figure 7: Figure was meant to include relative entropy (Kullback-Leibler divergence) of the model, as a fraction of the conditional histogram entropy. The values for the three histograms shown are: $\Delta H/H = 0.0292,\ \ 0.0212,\ \ 0.034$.
- Figure 9: Oriented subbands should be $\{B_0, B_1, \ldots, B_{K-1}\}$.
- Reference [8]: Correct reference is:
  A J Bell and T J Sejnowski. The 'Independent Components' of Natural Scenes are Edge Filters. *Vision Research*, 37(23):3327–3338, 1997.