Improved Denoising of Images Using Modelling of a Redundant Contourlet Transform

Boaz Matalon[a], Michael Elad[b] and Michael Zibulevsky[a]

[a]EE department, The Technion, Haifa 32000, Israel;
[b]CS department, The Technion, Haifa 32000, Israel

ABSTRACT

In this work we investigate the image denoising problem. One common approach found in the literature involves manipulating the coefficients in the transform domain, e.g. shrinkage, followed by the inverse transform. Several advanced methods that model the inter-coefficient dependencies were developed recently, and were shown to yield significant improvement. However, these methods operate on the transform domain error rather than on the image domain one. These errors are in general entirely different for redundant transforms. In this work we propose a novel denoising method, based on the Basis-Pursuit Denoising (BPDN). Our method combines the image domain error with the transform domain dependency structure, resulting in a general objective function, applicable for any wavelet-like transform. We focus here on the Contourlet Transform (CT) and on a redundant version of it, both relatively new transforms designed to sparsely represent images. The performance of our new method is compared favorably with the state-of-the-art method of Bayesian Least Squares Gaussian Scale Mixture (BLS-GSM), which we adapted to the CT as well, with further improvements still to come.

Keywords: denoising, sparsity, contourlet transform, redundancy, Gaussian scale mixtures, basis-pursuit, Laplacian distribution, inter-coefficient dependency

1. INTRODUCTION

In this work we focus on the problem of denoising images contaminated by additive white Gaussian noise. Symbolically, let x be the unknown clean image, n the additive noise and y the observed noisy image, i.e. $y = x+n$. Then denoising is defined as retrieving a reconstructed image $\hat{x}$, such that $\hat{x} \simeq x$ under some optimality criterion. Many recently developed denoising methods use the same basic algorithm: 1) Transform: calculate the coefficients of the given image in a chosen basis or frame 2) Operate: modify these coefficients in some way 3) Inverse transform: reconstruct the image. The most common way of such manipulation is shrinkage, namely performing a look-up-table (LUT) operation on each coefficient separately.[1, 2] Albeit simple, such approach ignores the inevitable inter-coefficient dependency. As we turn to use the more effective redundant transforms, this overlooked dependency further increases.

More advanced methods[3–5] try to model these dependencies, thus improving the performance while also complicating the algorithm. A drawback shared by these algorithms is their focus on removal of the noise in the transform domain, rather than in the image domain. Formulating optimality criteria with respect to the representation coefficients does not guarantee a successful treatment. Counter to the above algorithms, there exist several methods[6, 7] that relate the denoising objective directly to the image-domain error, and obtain the denoised image by minimization of a cost function. Nevertheless, their performance is often surpassed by the above transform-domain techniques.

In this paper we propose a new denoising method, built as a merge of these two distinct approaches. It minimizes an objective function containing the measurement error and a prior penalty. This penalty emerges from an approximate joint probability model for adjacent transform-domain coefficients, and thus can describe their inter-dependencies. This method can be easily extended to colored Gaussian noise, as well as to the

Further author information: (Send correspondence to B.M.)
B.M.: E-mail: mboaz17@tx.technion.ac.il
M.Z.: E-mail: mzib@ee.technion.ac.il
M.E.: E-mail: elad@cs.technion.ac.il

reconstruction of noisy and blurred images problem. Although we concentrate here on the contourlet transform and its redundant version (see below), this method is valid for any wavelet-like transform. In addition, we adapt the Gaussian-Scale-Mixture (GSM) model,[4] originally developed for steerable wavelets, to contourlets. We compare both denoising methods, showing that (a) taking into account coefficients dependencies is helpful; (b) redundancy improves the denoising results; and (c) the proposed approach leads to state-of-the-art performance, while being of manageable complexity and having clearer objective.

2. THE CONTOURLET TRANSFORM

It is well known that many signal processing tasks, e.g. compression, denoising, feature extraction and enhancement, benefit tremendously from having a parsimonious representation of the signal at hand. Do and Vetterli have conceived the Contourlet Transform[8] (CT), which is one of several transforms developed in recent years, aimed at improving the representation sparsity of images over the Wavelet Transform[9] (WT). The main feature of these transforms is the potential to efficiently handle 2-D singularities, i.e. edges, unlike wavelets which can deal with point singularities exclusively. This difference is caused by two main properties that the CT possess: 1) the *directionality* property, i.e. having basis functions at many directions, as opposed to only 3 directions of wavelets 2) the *anisotropy* property, meaning that the basis functions appear at various aspect ratios (depending on the scale), whereas wavelets are separable functions and thus their aspect ratio equals to 1. The main advantage of the CT over other geometrically-driven representations, e.g. curvelets[10] and bandelets,[11] is its relatively simple and efficient wavelet-like implementation using iterative filter banks. Due to its structural resemblance with the wavelet transform, many image processing tasks applied on wavelets can be seamlessly adapted to contourlets. Because of these features we chose to employ this transform throughout our work.

2.1. Basic Transform

The CT is constructed by two filter-bank stages, a Laplacian Pyramid[12] (LP) followed by a Directional Filter Bank[13] (DFB). The LP decomposes the image into octave radial-like frequency bands, while the DFB decomposes each LP detail band into many directions (a power of 2). Both of the stages are critically sampled, hence the transform is up to 33% redundant (due to the redundancy of the LP). In addition, the separability of the stages allows different number of directions for each radial band. Figure 1(a) shows a sample frequency partition of the CT, where three radial bands are divided into $8, 4$ and 4 directional subbands, from fine to coarse. This partition is not accidental – it has been proven[8] that minimal asymptotic approximation error is achieved when the number of directions is multiplied at every other finer scale. This concept, which appears also in the curvelet transform, as proposed by Candés and Donoho,[10] emerges from the optimal scaling law of $w \sim l^2$, where w and l symbolize the effective width and length of a contourlet. Figure 1(b) demonstrates this rule: when the width is divided by 4 (i.e. two scales finer), the length is divided by $\sqrt{4} = 2$.

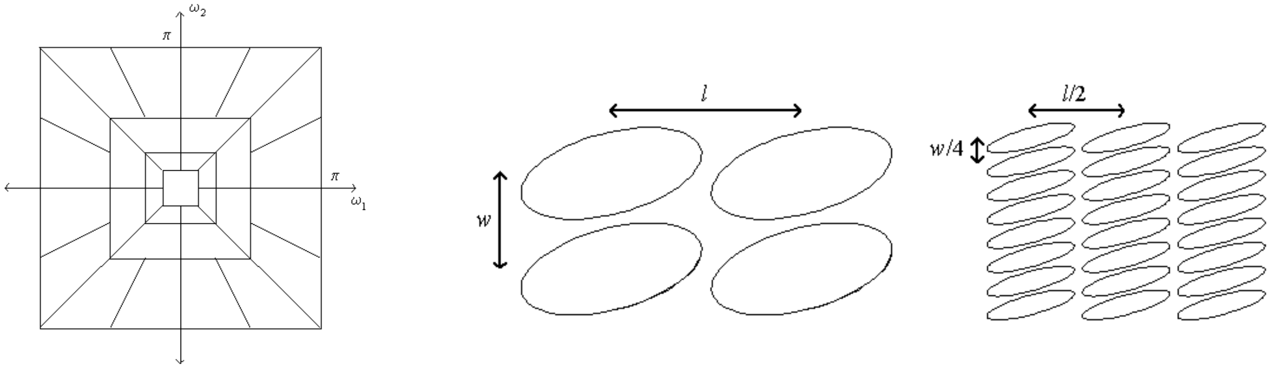


Figure 1. The CT: (a) Sample frequency partition. (b) Effective spatial support of subbands at two scales apart.

2.2. Extensions

The CT is a shift-variant transform, as it involves sampling at both the LP and the DFB stages. However, shift-variance is not a desirable property for various signal-processing tasks, and specifically denoising. To overcome this problem, a translation-invariant version of the CT was proposed by Cunha *et al.*, called the Nonsubsampled-Contourlet-Transform[14] (NSCT). This transform eliminates all sub-sampling operations, resulting in high redundancy, but accompanied by huge memory and computational requirements.

In this work the new denoising method was not employed on the NSCT, because of the enormous memory and running time requirements. Instead, a less-redundant version was used, which we call the Semi-Rotation-Invariant-Contourlet-Transform, or simply SRICT. As suggested by the name, only the DFB sub-sampling operations are eliminated, while the LP is still critically sampled. This means that at each finer scale, the amount of coefficients is multiplied by 4 if the number of directions stays the same, or by 8 if the number of directions is doubled. Therefore the total redundancy is

$$\left(\sum_j 2^{l_j}/4^{j-1}\right) - 1, \tag{1}$$

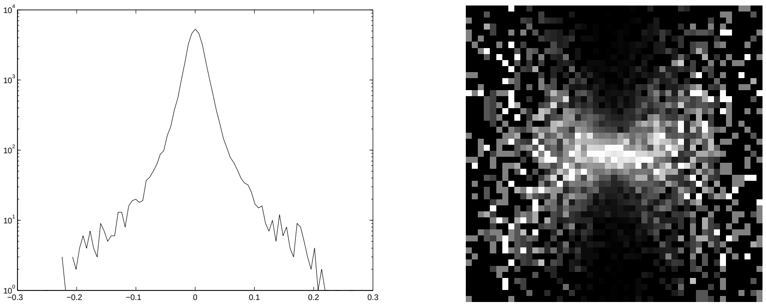
where 2^{l_j} denotes the number of directions at the j-th scale ($j = 1$ is the finest scale), and $l_j = 0$ for the remaining lowpass coefficients.

3. GAUSSIAN SCALE MIXTURE MODEL FOR CONTOURLETS

The Bayesian Least Squares Gaussian Scale Mixture (BLS-GSM) is a recently developed method for image denoising,[4] which achieves state-of-the-art results. It is based on statistical modelling of the coefficients of a multiscale oriented frame, specifically the Steerable Wavelet Transform,[15] but can be applied to other transforms as well. We will first describe briefly the method, and then elaborate on its application to the CT and the SRICT.

3.1. Description

It has been known for some time that images behave in a non-Gaussian fashion,[16] both at the image and the transform domain. This can be easily observed in the log marginal histogram of a bandpass filter response for some sample image, as shown in Fig. 2(a). The histogram is characterized by a *kurtotic* behavior, i.e. a sharp peak at zero, and tails that decay much slower than a Gaussian of the same variance.



(a) (b)

Figure 2. Histograms of one subband from the CT of *Peppers*: (a) log marginal; (b) conditional (each column has been separately rescaled to fit the display range).

The bandpass filter responses exhibit also non-Gaussian joint statistical behavior, not only marginal one. Specifically, coefficients at close spatial position, scale and orientation, show strong dependencies that cannot be vanished by decorrelation.[17] Firstly, large coefficients in a bandpass response of a natural image are mostly clustered together, which is particulary evident near edges. Secondly, the distribution of a coefficient conditioned by its neighbor value resembles a bow-tie shape (see Fig. 2(b)).

One way of describing both the marginal and the joint statistics of coefficients at the transform domain is by the Gaussian Scale Mixture (GSM) model.[4] In this model a local neighborhood is represented by a product of a Gaussian vector and an independent scalar multiplier. Formally, denote $\mathbf{z}$ as a local neighborhood of a reference coefficient, $\sqrt{\alpha}$ as a positive scalar multiplier and $\mathbf{u}$ as a zero-mean Gaussian vector. Then the basic GSM model assumption is

$$\mathbf{z} = \sqrt{\alpha}\mathbf{u}. \tag{2}$$

α is known as the *hidden multiplier*, since it cannot be observed directly. In other words, the vector $\mathbf{z}$ is an infinite mixture of Gaussian vectors, and its probability density function (pdf) is determined by the covariance matrix $\mathbf{C_u}$ of $\mathbf{u}$ and the pdf $p(\alpha)$. Many distributions can be represented as a GSM model (depending on $p(\alpha)$), e.g. the generalized Gaussian family and the symmetrized Gamma family.[18]

Since the GSM model specifies only a local description, the problem of obtaining a global model should be dealt with. One possibility is to divide the transform domain into non-overlapping neighborhoods, such that an independent GSM model is attached to each neighborhood.[18] Unfortunately, this simple idea results in noticeable artifacts at the neighborhood boundaries when performing denoising. Another option is to describe every local neighborhood by a GSM model, which means that each coefficient belong to many neighborhoods. This yields implicitly a global Markov model, which has a very complicated structure, thus making a task like denoising extremely difficult. The remaining possibility, which is the chosen one, is simply performing denoising to each coefficient according to its vicinity, ignoring the inevitable statistical dependency between overlapping neighborhoods. Note that the lowpass coefficients of the noisy image are not modified, since the noise is attenuated strongly there.

Following the above discussion, an observed noisy neighborhood $\mathbf{v}$ can be expressed as

$$\mathbf{v} = \sqrt{\alpha}\mathbf{u} + \mathbf{w}, \tag{3}$$

where $\mathbf{w}$ is the additive noise Gaussian vector, and all three random variables on the right side of Eq. (3) are independent. Assuming known noise characteristics at the image domain, the covariance matrix $\mathbf{C_w}$ of $\mathbf{w}$ can be calculated in advance. The covariance matrix $\mathbf{C_v}$ of $\mathbf{v}$ is presumed constant for all of the coefficients in a certain bandpass, and is thus estimated by the sample covariance. If we calculate now $E\{\mathbf{v}\mathbf{v}^t\}$, we get $\mathbf{C_v} = E\{\alpha\}\mathbf{C_u} + \mathbf{C_w}$. Without loss of generality, one can assume $E\{\alpha\} = 1$, and thus we finally get $\mathbf{C_u} = \mathbf{C_v} - \mathbf{C_w}$.

Although the multiplier's distribution $p(\alpha)$ can be estimated from the observed data by a Maximum Likelihood approach, the best PSNR results were obtained by using Jeffrey's prior,[19] which in this case reduces to

$$p(\alpha) \propto \frac{1}{\alpha}. \tag{4}$$

Since this is not a proper pdf, practically we set $p(\alpha) = 0$ outside $[\alpha_{min}, \alpha_{max}]$, where α_{min} and α_{max} are a small and a large positive constant, respectively.

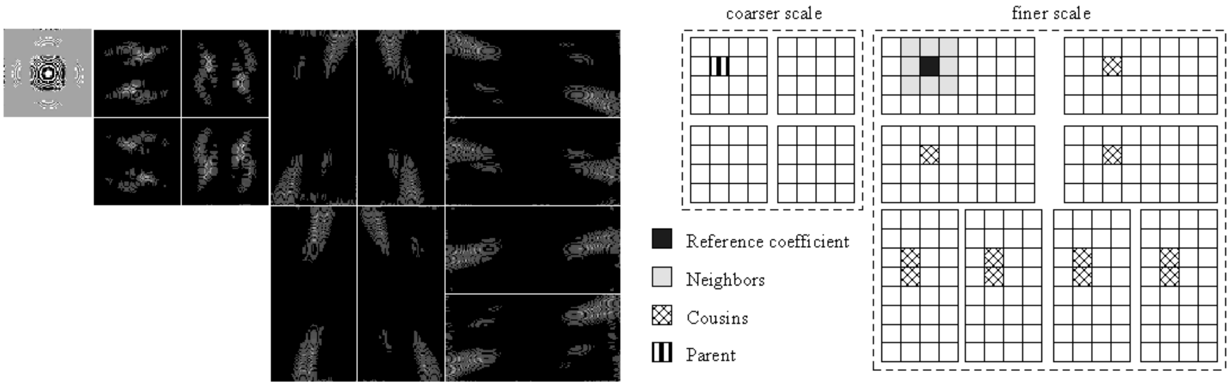
Now that all of the variables' distributions are known (or estimated), we can write the Bayes Least Squares (BLS) estimate of the reference coefficient z_c from the observed neighborhood $\mathbf{v}$. After some mathematical manipulations, it is given by

$$E\{z_c|\mathbf{v}\} = \int_0^\infty p(\alpha|\mathbf{v})E\{z_c|\mathbf{v},\alpha\}d\alpha. \tag{5}$$

The factor $E\{z_c|\mathbf{v},\alpha\}$ is simply a local Wiener estimate, since $\mathbf{z}$ and $\mathbf{v}$ are Gaussian for a given α. Hence, this formula can be interpreted as an infinite weighted average of Wiener estimates. In practice, the integration is replaced by a finite summation over many multiplier's values. After all of the coefficients are modified via Eq. (5), the image is reconstructed by the inverse transform. As mentioned earlier (Sect. 1), all of these manipulations take place in the transform domain, and thus while there is clearly a relation to the actual image domain noise, this relation is not entirely accurate and exhaustive.

3.2. Application to the CT and the SRICT

Once the neighborhood is defined for a certain representation, the BLS-GSM method can be employed. To specify a meaningful neighborhood, we need to look first at the structure of the CT. Figure 3(a) shows the contourlet coefficients of the image *Zoneplate*, for a certain frequency partition (4 and 8 directions). It is readily apparent that the coefficients are organized in a quadtree, where each coefficient has four children at the immediate finer scale, at the same spatial location, and either at the same direction or at two finer directions (when the number of directions is doubled). Figure 3(b) shows the nominees to belong to a neighborhood of a particular coefficient[3]: 1) the parent, i.e. the coefficient in the same spatial location at the immediate coarser scale. 2) the neighbors, i.e. the eight adjacent coefficients at the same subband. 3) the cousins, i.e. the coefficients at the same scale and spatial location, but different direction. In our experiments, the best results were obtained with neighborhoods that include only a parent and the 8 nearest neighbors, and this choice will be referred to hereafter.



(a) (b)

Figure 3. The CT: (a) coefficients of *Zoneplate*; (b) coefficient relationships

The neighborhood definition of the SRICT is a little bit trickier than this of the CT. Since the SRICT is much more redundant, the effective supports of basis functions of adjacent coefficients at the same subband coincide. Thus, choosing the same neighborhood as before results in very correlative neighbors. Another option is to take neighbors that their respective basis functions do not coincide (effectively), and thus the correlation between neighbors is very small. In our experiments the latter option yielded slightly better results, and therefore was chosen (with 8 spatial neighbors). As a side note, notice that a parent has eight children when the number of directions is doubled, rather than four children in the CT.

4. ALTERNATIVE DENOISING METHOD

This section describes a novel method for image denoising, which is basically minimization of a cost function, incorporating a new global image model. As opposed to recently developed methods, this approach refers to the image domain error, rather than the transform domain error. Since minimization of the MSE at the transform domain does not translate directly to MMSE at the image domain for non-orthonormal transforms, a fundamental flaw lies within many state-of-the-art methods, like the BLS-GSM.

4.1. Formulation

Let us turn our attention first to the Basis-Pursuit De-Noising (BPDN) method, which was introduced by Chen, Donoho and Saunders.[6] It refers to the solution of

$$\hat{\mathbf{z}} = \min_{\mathbf{z}} \; \frac{1}{2}\|y - \Phi\mathbf{z}\|_2^2 + \lambda\|\mathbf{z}\|_1, \tag{6}$$

where Φ represents the *synthesis* transform operator, $\mathbf{z}$ the coefficients vector, and λ an adjustable parameter. The reconstructed image is given by $\hat{x} = \Phi\hat{\mathbf{z}}$. This is essentially the maximum a-posteriori probability (MAP) solution of the denoising problem, where the transform coefficients are modelled as independent *Laplacian* random variables. More specifically, each coefficient is distributed according to $p(z) \propto \exp(-\frac{\sqrt{2}}{\sigma}|z|)$, where σ is the standard deviation.

This objective function can be generalized somewhat by allowing each coefficient z_i to have its own weight λ_i, and thus we get

$$\hat{\mathbf{z}} = \min_{\mathbf{z}} \frac{1}{2}\|y - \Phi\mathbf{z}\|_2^2 + \sum_i \lambda_i |z_i|. \tag{7}$$

With respect to a multiscale transform, such as the Contourlet transform, experiments made on natural images show that coefficients at different scales and directions have different average standard deviation. Hence σ should depend on the scale and direction, and perhaps the spatial position as well, which justifies a coefficient dependent weight λ_i, as indicated above.

A possible downside of such an approach is the statistical independence assumption of different coefficients. As later results will show, this approach is inferior to the proposed methods, which explicitly model inter-coefficient dependencies. In developing these methods, the main challenge arising is how to formulate a global prior model from the local ones described earlier. However, we must emphasize that these local models serve only as an intuition, since they correspond to the analysis operator response, not necessarily to the underlying distribution.

We saw in Sect. 3.1 that the empirical distribution of a local neighborhood of a coefficient can be quite accurately described by a Gaussian Scale Mixture (GSM) model. However, an analytic expression for this distribution cannot be obtained, thus ruling out its use as a prior.

Sendur and Selesnick[5] suggested the use of a new bivariate pdf to model the distribution of a coefficient and its parent. They employed this pdf to construct a MAP-based *bivariate* shrinkage rule, unlike the commonly used *scalar* shrinkage rules. In contrast with the GSM model, this new model has a simple analytic form, yet it still retains good approximation of the empirical distribution. More specifically, the joint pdf of a coefficient z_1 and its parent z_2 is given by

$$p(z_1, z_2) = \frac{3}{2\pi\sigma_1\sigma_2} \exp\left(-\sqrt{3}\sqrt{\left(\frac{z_1}{\sigma_1}\right)^2 + \left(\frac{z_2}{\sigma_2}\right)^2}\right), \tag{8}$$

where σ_i corresponds to the standard deviation of z_i.

This model can be easily extended to account for the dependencies in a local neighborhood with arbitrary size. Denote $\mathbf{z} = (z_1, z_2, \ldots, z_n)$, where z_j is the j-th coefficient in the neighborhood (z_1 is the central coefficient). In addition, denote σ_j as the standard deviation of z_j. Then the joint pdf is given by

$$p(\mathbf{z}) = K \exp\left(-a\sqrt{\sum_j \left(\frac{z_j}{\sigma_j}\right)^2}\right), \tag{9}$$

where K is a normalizing factor, and a ensures that σ_j is indeed the standard deviation of z_j.

To examine the accuracy of the model of Eq. (9), it can be compared with an empirical histogram. Figure 4(a) shows the log joint histogram of a reference coefficient and one of its nearest neighbors, estimated from the finest CT bands of several images. The values in each band were first scaled down by the subband's standard deviation, to get a single distribution rather than a mix of distributions. Two main deviations from the discussed model can be easily observed in the empirical histogram: 1) The model suggests non-smooth surface for $\mathbf{z} = 0$, but it is in fact smooth. 2) The decay rate diminishes as $|\mathbf{z}|$ increases, while the model suggests a constant decay rate. These phenomena are seen in Fig. 2(a) as well, for a 1-D pdf.

The first difference can be solved by adding a small positive constant ε into the square root of Eq. (9), thus making the surface smooth near the origin. To overcome the second difference, the model should impose concaveness for large values of $|\mathbf{z}|$. The simplest way of achieving it is by decreasing the power inside the exponent from $1/2$ to $1/\gamma \quad (\gamma > 2)$. Thus, the modified model is given by

$$p(\mathbf{z}) = K \exp\left(-a\left(\sum_j \left(\frac{z_j}{\sigma_j}\right)^2 + \varepsilon\right)^{\frac{1}{\gamma}}\right). \tag{10}$$

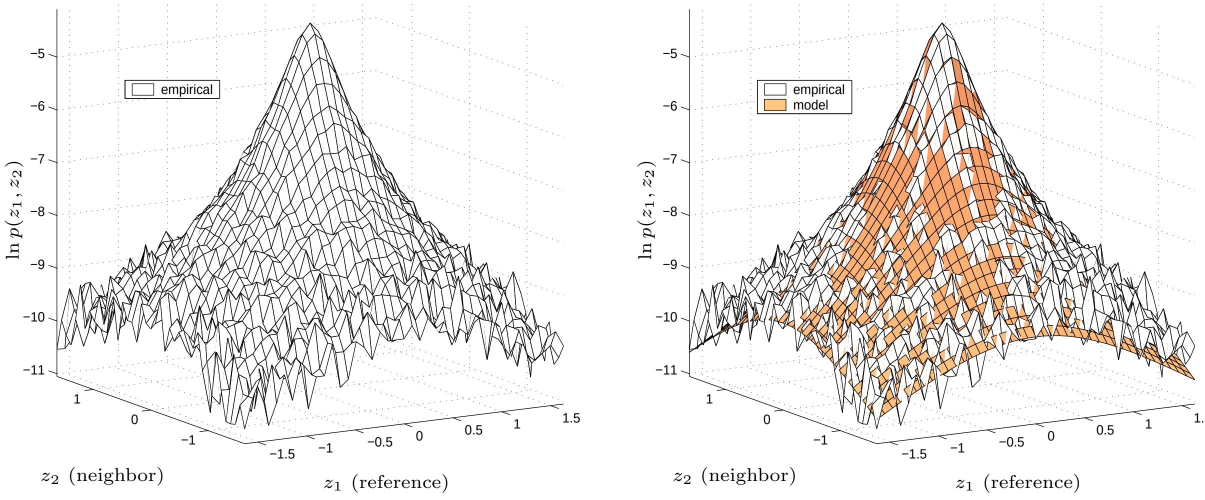


Figure 4. Log joint histogram of two nearest neighbors: (a) empirical; (b) empirical vs. the proposed model in (10).

Figure 4(b) depicts both the estimated and the modelled log joint pdf, manually optimized to obtain visual fitting (optimal values are $\varepsilon = 2.5 \cdot 10^{-2}, \gamma = 6$).

This model can be further generalized by introducing correlations between coefficients. Although the current pdf is not separable, it is certainly symmetric with respect to every axis, and thus all of the coefficients are uncorrelated. This assumption may be valid for a small-redundancy transform such as the CT, as seen in figure 2(b), but is completely invalid for a redundant transform, where the correlation increases with the redundancy. To overcome this problem, the previous model can be modified to include a covariance matrix C, such that

$$p(\mathbf{z}) = K \exp\left(-a\left(\mathbf{z}C^{-1}\mathbf{z} + \varepsilon\right)^{\frac{1}{\gamma}}\right). \tag{11}$$

Note that Eq. (11) reduces to Eq. (10) by imposing a diagonal covariance matrix.

The question arising now is how to extend the local prior model into a global one. As discussed in Sect. 3.1, the local GSM model is implicitly extended to a global model, by assuming independent scaling variables for different neighborhoods, even for overlapping ones. Remember this was done because of the great difficulty to perform statistical inference with a complicated global dependency model. Sendur and Selesnick[5] also assumed independent neighborhoods, in order to achieve a relatively simple shrinkage rule.

In a similar fashion, we also embrace the independency assumption, so that the global pdf is given by

$$p(\mathbf{z}) \propto \prod_i \exp\left(-a\left(\mathbf{z}_i C_i^{-1}\mathbf{z}_i + \varepsilon\right)^{\frac{1}{\gamma}}\right). \tag{12}$$

Here we denote $\mathbf{z}$ as the global coefficients vector, $\mathbf{z}_i$ as the i-th neighborhood coefficients vector and C_i as the corresponding covariance matrix. Replacing the independent Laplacian prior with the latter prior, Eq. (7) becomes

$$\hat{\mathbf{z}} = \min_z \frac{1}{2}\|y - \Phi\mathbf{z}\|_2^2 + \lambda \sum_i \left(\mathbf{z}_i C_i^{-1}\mathbf{z}_i + \varepsilon\right)^{\frac{1}{\gamma}}, \tag{13}$$

where λ is again an adjustable constant. Note that the summation in Eq. (13) is not made over the lowpass coefficients, since the discussed dependency model is not valid for these. This method will be denoted hereafter BPDN-COV.

Using the slightly less general model of Eq. (10) instead of Eq. (11), another version of the new denoising method, which will be denoted by BPDN-VAR, admits the form

$$\hat{\mathbf{z}} = \min_{z} \frac{1}{2}\|y - \Phi\mathbf{z}\|_2^2 + \lambda \sum_{i} \Big(\sum_{j(i)} \Big(\frac{z_j}{\sigma_j} \Big)^2 + \varepsilon \Big)^{\frac{1}{\gamma}}, \tag{14}$$

where $\{j(i)\}$ indicates the indices of the coefficients included in the i-th neighborhood. Notice that σ does not depend on i, thus the value of σ_j is shared between all of the neighborhoods z_j belongs to. On the other hand, in the BPDN-COV method each neighborhood has its own covariance matrix, hence a coefficient may be attached with several different variances.

4.2. Variance Estimation

The implementation of the new algorithm requires an estimation of either the variances $\{\sigma_i\}$ (BPDN-VAR method) or the covariances $\{C_i\}$ (BPDN-COV method) from the given data. Due to the model's complexity, a structured estimator like the maximum-likelihood (ML) estimator is very difficult to develop, thus an heuristic estimator must be obtained. One common way of estimating the variances (as in Ref. 20) is by using coefficients from the reference coefficient's vicinity. Denote $\sigma_{n,i}^2$ as the noise variance at the i-th coefficient (it is in fact constant over a subband). Then the variance estimate of the i-th coefficient is given by

$$\hat{\sigma}_i^2 = \max \Big\{ \frac{1}{\#j} \sum_{j(i)} z_j^2 - \sigma_{n,i}^2 \quad ,0 \Big\}, \tag{15}$$

where $\#j$ signifies the number of coefficients in the local neighborhood. Although this estimate makes sense, it is too sensitive to the neighborhood's size: a small one leads to an unreliable estimation, while a large one yields slow adaptation to varying characteristics. As a result, the reconstructed images in our experiments obtained in this method were blotchy, and therefore this method was abandoned.

An alternative estimation method was introduced by Chang *et al.*[21] for the WT, though it remains valid for any multiscale transform like the CT. Consider a subband with M coefficients, and denote $\bar{\mathbf{z}}_i$ as a $p \times 1$ vector containing the *absolute values* of p neighbors of z_i. The *context* of z_i is defined as a weighted average of its neighbors' absolute values $y_i = \mathbf{w}^t \bar{\mathbf{z}}_i$. The weights vector $\mathbf{w}$ is calculated by the least squares (LS) estimate over the whole subband, i.e.

$$\mathbf{w}_{LS} = (Z^t Z)^{-1} Z^t |\mathbf{z}|, \tag{16}$$

where Z is a $M \times p$ matrix with rows $\{\bar{\mathbf{z}}_i\}$, and $\mathbf{z}$ is a $M \times 1$ vector of the subband's coefficients. As shown in Fig. 4, the contourlet coefficients are essentially uncorrelated, and thus a linear predictor based on their values cannot carry much information about another coefficient. Yet, the absolute values of neighbors are correlated,[22] which explains their use in Eq. (16).

Following the context calculation, a coefficients' variance is estimated based on all of the coefficients in the subband with similar context. More precisely, the contexts $\{y_j\}$ are sorted in an increasing order, and the coefficients $\{z_j\}$ whose context are at most L values away from y_i are chosen (i.e. $2L + 1$ coefficients). The variance estimate of z_i is then

$$\hat{\sigma}_i^2 = \max \Big\{ \frac{1}{2L+1} \sum_{j(i)} z_j^2 - \sigma_{n,i}^2 \quad ,0 \Big\}. \tag{17}$$

As Fig. 2(b) demonstrates, a coefficients' standard deviation scales roughly linearly with its neighbor's absolute value. Hence, the above method can be understood as gathering of coefficients with the same variance, then estimating this variance. Similarly to Ref. 21, we choose $L = \max\{100, 0.02M\}$ to guarantee reliable estimation along with adaptivity to varying characteristics. In addition, we set $p = 9$ (eight spatial neighbors and one parent), yet other values ($p = 5, 8$) led to the same performance.

In a similar fashion to the extension of this method for the shift-invariant WT,[21] we now extend it for the SRICT. Since the estimator of Eq. (17) originates from the ML estimator in the Gaussian i.i.d. case, it will be unreliable if the data samples are highly correlative, which is the case for adjacent coefficients of the SRICT. To prevent such a scenario, a subband is first partitioned into several clusters with little intra correlation. This is done by grouping together of basis functions that their effective supports do not coincide (see Fig. 1(b)), and thus a subband in a 2^l-directional scale is divided into 2^l clusters. From this point each cluster is treated as a separate subband with respect to Eqs. (16) and (17), hence the correlation problem is circumvented.

Estimation of the covariances in the BPDN-COV method is a more challenging task than the variances estimation. A simple approach utilized in the BLS-GSM method (Sect. 3.1) contains imposing a single covariance matrix to each subband, and estimating it by the sample covariance. Although this approach yields reliable estimation due to the small number of unknown parameters, it fails to adapt to spatially changing characteristics. As a consequence, the denoising performance of this method in our experiments was quite poor.

To allow spatial adaptation while still retaining reliability, we propose a variation of the context-driven variance estimation method described earlier. Maintaining the same notations, the estimate of C_i is given by

$$\hat{C}_i = \frac{1}{2L+1} \sum_{j(i)} \mathbf{z}_i \mathbf{z}_i^t - C_{n,i}, \tag{18}$$

where $C_{n,i}$ represents the noise covariance at the i-th neighborhood (once again, it depends on the subband only). To assure that $\hat{C}_i \succ 0$, its eigen-values are increased to become positive, if necessary. The idea behind this approach is that coefficients with similar variances, also have associated neighborhoods with similar covariances. However, the enormous amount of unknown parameters makes this suggestion impractical.

To significantly reduce the number of parameters, a compromise between the two above approaches is made. The subband's coefficients are uniformly classified into several groups (for example 50), based on their context, and each group is attached with a single covariance matrix. Then the estimator of Eq. (18) is obtained, where the reference context value of a group is its median. Of course, since a covariance matrix contains many unknown parameters, more data values are needed than in the variance estimation case, and thus we choose $L = \max\{750, 0.05M\}$. The results of this approach was substantially better PSNR-wise than those of the non-adaptive covariance approach. However, its computational complexity still proved too expensive, which prevented its thorough examination on large images. We thus leave this matter for future work and concentrate on the BPDN-VAR method.

5. EXPERIMENTS

5.1. Implementation Issues

The BPDN-VAR method contains several unknown parameters which must be selected: $\gamma, \varepsilon, \lambda$ and the neighborhood size. As discussed earlier (Sect. 4.1), γ and ε can be set manually to fit the 2-D joint histogram (see Fig. 4). However, such a choice might not be suitable for higher dimensional distributions. Moreover, for $\gamma > 2$ the objective function in Eq. (14) is not convex, necessitating a sequential minimization for increasing values of γ. Therefore, in this paper we set $\gamma = 2$, although other values will be examined in a future work. The value of ε must be positive to ensure a smooth objective function, and also to allow better fitting of the empirical histogram to the model. On these grounds and based on our experiments we have chosen $\varepsilon = 10^{-2}$.

Regarding the neighborhood selection, it has been shown[3] that the dependency between a coefficient and its parent exceeds that of any other neighbor. Thus the parent is included in all of the neighborhoods, other than those at the coarsest scale, because no parent exists there. For the CT, the choice of spatial neighbors which led to the best performance was of the four nearest neighbors, unlike eight neighbors in the GSM-BLS method (see Sect. 3.2). The same stands for the SRICT, only that we refer to the neighbors whose supports do not coincide. For comparison, we will also examine the 1×1 neighborhood case (i.e. the reference coefficient alone), which will be denoted by BPDN-VAR-NN (stands for No-Neighbors).

Returning briefly to Eq. (7), and remembering that it corresponds to the MAP-solution for an independent Laplacian prior model, we get $\lambda_i = \sqrt{2}\sigma_n^2/\sigma_i$, where σ_n^2 is the noise variance at the image domain. Going back to the BPDN-VAR-NN method, the corresponding value of λ is $\lambda_0 = \sqrt{2}\sigma_n^2$, which turned out to be indeed the optimal value performance-wise. However, in the BPDN-VAR method (see Eq. (14)), each coefficient appears five times in the summation if it belongs to the finest scale, or nine times otherwise. Clearly no value of λ exists such that the 'effective' weight of each coefficient equals $\sqrt{2}\sigma_n^2/\sigma_i$. One possible solution is to multiply σ_i^2 by $(9/5)^2$, except for coefficients at the finest scale, and also to set $\lambda = \lambda_0/5$. For the CT, the same PSNR values were reached for $\lambda \in [\lambda_0/5, \lambda_0/2]$, yet better visual quality was seen by setting $\lambda = \lambda_0/2$, hence this value was finally chosen (likewise, $\lambda = 10\lambda_0/7$ proved optimal for the BPDN-VAR-NN method). This choice indeed exceeded all other possibilities performance-wise (e.g. a certain λ without changing σ_i^2). For the SRICT, a slightly different manipulation of σ_i^2 was made,[23] and $\lambda = \lambda_0/5$ yielded the best results (or $\lambda = \lambda_0/3$ for BPDN-VAR-NN). As said before, the PSNR values change very little with λ, proving the robustness of our method.

Following the selection and estimation of the unknown parameters, the minimization of the cost function in Eq. (14), denoted hereafter by $f(\mathbf{z})$, can begin. It must be emphasized that rather than obtaining an accurate solution of (14), our goal is to reach the highest quality image fast. A work by Elad[2] showed that the BPDN problem (Eq. (6)) can be solved by iteratively performing simple shrinkage on the coefficients. This work can be easily extended to apply on the BPDN-VAR-NN method by making a certain modification to the coefficient-dependent thresholds. In our experiments, this technique produced satisfactory results much faster than any other optimization technique. Nevertheless, the BPDN-VAR method cannot be expressed as a series of closed-form LUT operations. This distinction vastly increases the complexity of the discussed technique, thus ruling out its use for BPDN-VAR method.

After testing many optimization algorithms, we finally decided to use the Truncated-Newton algorithm with preconditioning[24, 25] for BPDN-VAR method. In short, the current solution $\mathbf{z}_k$ is updated by $\mathbf{z}_{k+1} = \mathbf{z}_k + \alpha_k \mathbf{d}_k$, where α_k is obtained from a line-search along the direction $\mathbf{d}_k$. This direction is an approximate solution of $H_k\mathbf{d} = -\mathbf{g}_k$, where H_k and $\mathbf{g}_k$ are the hessian and the gradient of $f(\mathbf{z})$ at $\mathbf{z}_k$, respectively. Such an approximation can be made by the Conjugate-Gradients method, sped up by employing a diagonal preconditioner. More details about the optimization method, including explicit expressions for the gradient and the hessian's diagonal, can be found in Ref. 23.

A major difference in the PSNR behavior between the CT and the SRICT was revealed in our experiments. As for the CT, the PSNR increased with every multidimensional iteration, until it settled after about 15 iterations. On the other hand, for the SRICT there was a quick rise in the PSNR, then about 4 iterations of peak value, followed by a small decrease until settling. This suggests that minimization of $f(\mathbf{z})$ does not translate to maximization of the PSNR for the SRICT. However, it does mean that premature termination is possible and even recommended. In all of the simulations the maximal PSNR value was reached as soon as 6 (for 256×256 images) or 7 (for 512×512 images) iterations, irrespective of the specific image at hand, and thus the BPDN-VAR method is well-defined.

To further assess the performance of our new methods, we have implemented two more algorithms. One is hard-thresholding (HT), namely zero-forcing z_i if it is smaller than a threshold $K\sigma_{n,i}$ (see Sect. 4.2 for notations). As in Ref. 14, we set $K = 4$ for the finest scale, and $K = 3$ otherwise. The other algorithm, which will be named as adaptive-soft-thresholding (AST), performs soft-shrinkage with a threshold of $\sigma_{n,i}^2/\hat{\sigma}_i$, where $\{\hat{\sigma}_i\}$ are calculated as in Sect. 4.2. This is the algorithm proposed by Chang *et al.*,[21] adapted to the CT.

5.2. Results

Figure 5 displays a 128×128 slice of *Peppers*, and its denoising results with the discussed methods, using the CT. The corresponding PSNR values appear in Table 1. The better visual quality of both the BPDN methods relative to standard shrinkage techniques is readily seen. In addition, although the PSNR of BLS-GSM is slightly higher, the BPDN-VAR method produces the same, if not better, visual quality.

Table 1 summarizes the PSNR results of all of the examined methods, for $\sigma_n = 20$. The test images were downloaded from http://decsai.ugr.es/˜javier/denoise. This comparison reveals some interesting observations: 1) Not surprisingly, adding redundancy improves the performance, although not dramatically. 2) The BPDN-VAR method surpasses the BPDN-VAR-NN method uniformly (on average, $0.24dB$ for CT, $0.28dB$ for SRICT). Therefore, modifying the prior to account for the dependencies is worthwhile. 3) For the CT, the BPDN-VAR method attains essentially the same PSNR as BLS-GSM (merely $0.02dB$ less). For the SRICT, the difference is larger ($0.11dB$ less). Yet, remember that the strong inter-coefficient correlations in the SRICT are handled by the BLS-GSM method, in contrast with BPDN-VAR. Thus, a more fair comparison, which is left to future work, would include the BPDN-COV method.

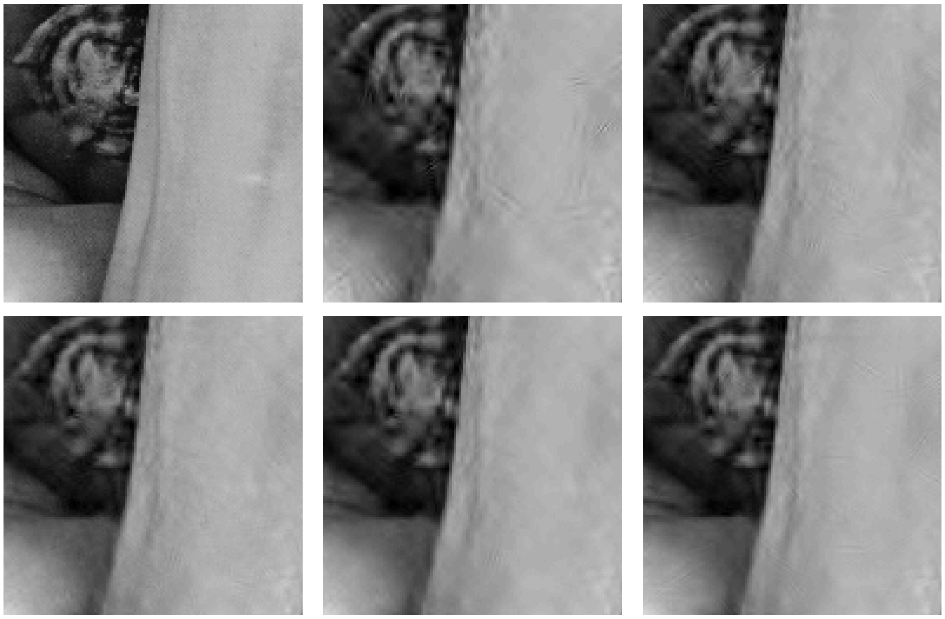


Figure 5. Denoising results of a 128×128 slice of *Peppers* (using the CT and for $\sigma_n = 20$). From left to right and top to bottom: Original; HT; AST; BPDN-VAR-NN; BPDN-VAR; BLS-GSM.

Table 1. PSNR values for all of the images and methods ($\sigma_n = 20$)

	Peppers256		*Peppers*		*Lena*		*Barbara*	
	CT	SRICT	CT	SRICT	CT	SRICT	CT	SRICT
HT	26.73	28.07	29.08	30.31	29.27	30.59	26.16	27.54
AST	28.59	29.05	30.52	30.79	30.91	31.14	28.66	29.08
BPDN-VAR-NN	28.81	28.99	30.79	30.92	31.19	31.23	28.92	28.81
BPDN-VAR	**29.02**	**29.25**	30.96	31.08	31.47	31.52	29.20	29.21
BLS-GSM	28.91	29.12	**31.08**	**31.30**	**31.49**	**31.62**	**29.26**	**29.44**

6. CONCLUSIONS

We have proposed a novel denoising method, by merging the inherent transform domain inter-coefficient dependencies into a MAP framework. The resulting algorithm proved competitive with the state-of-the-art method of BLS-GSM, which we adapted for the contourlet transform. The new approach is advantageous both in its direct reference to the image domain error, and its straight-forward extension to other inverse problems. It is also applicable to any wavelet-like transform.

There is still much work to be carried out concerning this method. First, the performance of the BPDN-COV method should be thoroughly investigated, while ways of speeding-up the optimization process need to be sought after. Also, various values of γ and ε (see Sect. 4.1) should be tested. Secondly, the same prior developed here could serve for more general inverse problems such as deblurring. As a last point we mention that additional study can be done to further modify the priors in Eqs. (13) and (14), in order to better describe the inter-coefficient dependencies. These and other directions are part of our future work plan.

REFERENCES

1. D. L. Donoho, "De-noising by soft thresholding," *IEEE Trans. on Information Theory* **41**, pp. 613–627, May 1995.
2. M. Elad, "Why simple shrinkage is still relevant for redundant representations?," *IEEE Trans. on Information Theory* , submitted, Jan. 2005.

3. D. D. Y. Po and M. N. Do, "Directional multiscale modeling of images using the contourlet transform," *IEEE Trans. on Image Processing* , to appear, 2005.
4. J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *IEEE Trans. on Image Processing* **12**, pp. 1338–1351, Nov. 2003.
5. L. Sendur and I. W. Selesnick, "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency," *IEEE Trans. Signal Processing* **50**, pp. 2744–2755, Nov. 2002.
6. S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing* **20**(1), pp. 33–61, 1999.
7. F. Dibos and G. Koepfler, "Global total variation minimization," *SIAM Journal on Numerical Analysis* **37**(2), pp. 646–664, 2000.
8. M. N. Do and M. Vetterli, "The contourlet transform: An efficient directional multiresolution image representation," *IEEE Trans. on Image Processing* , to appear, 2004.
9. S. Mallat, *A Wavelet Tour of Signal Processing*, Accademic Press, 2 ed., 1999.
10. E. J. Candés and D. Donoho, "New tight frames of curvelets and optimal representations of objects with smooth singularities," tech. rep., 2002.
11. E. L. Pennec and S. Mallat, "Sparse geometric image representation with bandelets," *IEEE Trans. on Image Processing* **14**, pp. 423–438, April 2005.
12. M. N. Do and M. Vetterli, "Framing pyramids," *IEEE Trans. on Signal Processing* **51**, pp. 2329–2342, Sep. 2003.
13. R. H. Bamberger and M. J. T. Smith, "A filter bank for the directional decomposition of images: Theory and design," *IEEE Trans. on Signal Processing* **40**, pp. 882–893, April 1992.
14. A. L. Cunha, J. Zhou, and M. N. Do, "The nonsubsampled contourlet transform: Theory, design, and applications," *IEEE Trans. on Image Processing* , submitted, 2005.
15. E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans. on Tnformation Theory* **38**(2), pp. 587–607, 1992.
16. S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **11**(7), pp. 674–693, 1989.
17. R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," Tech. Rep. 414, Munich, Germany, 1997.
18. M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in modeling and analyzing natural imagery," *Applied and Computational Harmonic Analysis* **11**, pp. 89–123, July 2001.
19. G. E. P. Box and C. Tiao, *Bayesian Inference in Statistical Analysis*, Addison Wesley, Reading, MA, 1992.
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," in *Proceedings DCC'97 (IEEE Data Compression Conference)*, March 1997.
21. S. G. C. B. Yu and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," *IEEE Trans. on Image Processing* **9**, pp. 1522–1531, Sep. 2000.
22. J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. on Signal Processing* **41**, pp. 3445–3462, Dec. 1992.
23. B. Matalon, M. Zibulevsky, and M. Elad, "A new method for image denoising," tech. rep., to appear, 2005.
24. S. G. Nash, "A survey of truncated-newton methods," *Journal of Computational and Applied Mathematics* **124**, pp. 45–59, 2000.
25. P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, New York, 1981.