IMAGE DENOISING WITH THE CONTOURLET TRANSFORM

Boaz Matalon, Michael Elad and Michael Zibulevsky

The Technion, Haifa, Israel

ABSTRACT

In this work the image denoising problem is examined. A common approach involves transform-domain coefficients manipulation, followed by the inverse transform. This approach is highlighted by recently-developed methods that model the inter-coefficient dependencies. However, these methods operate on the transform domain error rather than on the more relevant image domain one. In this work we propose a novel denoising method, based on the Basis-Pursuit Denoising (BPDN) method. 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), a relatively new transform designed to sparsely represent images. The superiority of our method over BPDN is demonstrated, thus providing a more advanced tool for image restoration.

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 $\hat{x} \simeq x$. Many recently developed denoising methods operate by manipulating the transform coefficients of the given image. The most common way of such manipulation is shrinkage, namely performing a look-up-table (LUT) operation on each coefficient separately [1]. 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 [2, 3, 4] 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, which does not guarantee a successful treatment. Counter to the above algorithms, there exist several methods [5, 6] 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. It is in fact a generalization of the Basis-Pursuit prior [5], which was also employed in this work for comparison. This novel method can be easily extended to colored Gaussian noise, as well as to the reconstruction of noisy and blurred images problem. Although we concentrate here on the contourlet transform (see below), this method is valid for any wavelet-like transform. In addition, we adapt the Gaussian-Scale-Mixture (GSM) model [3], originally developed for steerable wavelets, to contourlets. By comparing the above mentioned methods, we show that (a) taking into account coefficients dependencies is helpful; and (b) the proposed approach leads to state-of-the-art performance (for a given transform), 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 [7] (CT), which is one of several transforms developed in recent years, aimed at improving the representation sparsity of images over the Wavelet Transform (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 [8], 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.

In our previous work [9], the original CT was employed, as well as a much more redundant version of it. However, a new version of contourlets, called the Contourlet-SD [10], was recently supplied to us. This representation is only up to 133% redundant, but nevertheless produces considerably better results than both transforms previously used. Therefore, throughout this work we employ this new transform, which will still be denoted by CT.

3. GAUSSIAN SCALE MIXTURE MODEL FOR CONTOURLETS

The Bayesian Least Squares Gaussian Scale Mixture (BLS-GSM) is a recently developed method for image denoising [3], 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, but can be applied to other transforms as well. We will first describe briefly the method, then elaborate on its application to the CT.

3.1. Description

It has been known for some time that images behave in a non Gaussian fashion, both at the image and the transform domain. This can be easily observed in the log marginal histogram of a bandpass filter response for a sample image, as shown in Fig. 1 (left). The histogram is typical of a *kurtotic* behavior, i.e. a sharp peak at zero, and tails that decay much slower than a Gaussian of the same variance.

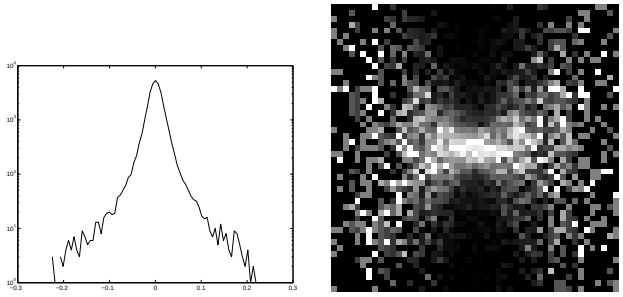

Fig. 1. Histograms of one subband from the CT of *Peppers* (left to right): log marginal; conditional (each column has been separately rescaled to fit the display range).

The bandpass filter responses exhibit also non-Gaussian joint statistical behavior, not only marginal one. Specifically, coefficients at close spatial positions, scales and orientations, show strong dependencies that cannot be vanished by decorrelation. 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. 1, right).

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 [3]. 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 $\mathbf{z} = \sqrt{\alpha}\mathbf{u}$, where α is known as the *hidden multiplier*, since it cannot be observed.

An observed noisy neighborhood $\mathbf{v}$ can be expressed as $\mathbf{v} = \sqrt{\alpha}\mathbf{u} + \mathbf{w}$, where $\mathbf{w}$ is the additive noise Gaussian vector, and all three random variables on the right side of the equation are independent. The reference coefficient is reconstructed by the Bayes Least Squares (BLS) estimate, given by

$$E\{z_c|\mathbf{v}\} = \int_0^\infty p(\alpha|\mathbf{v})E\{z_c|\mathbf{v},\alpha\}d\alpha. \quad (1)$$

After all of the coefficients are modified via Eq. (1), the image is reconstructed by the inverse transform.

3.2. Application to the CT

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 new CT [10]. Each of the two finest scales contains the same number of coefficients as the image, and thereafter the number of coefficients is divided by four every coarser scale. Independently, the number of directions is doubled every other finer scale. Hence, there are four possible parent-child relationships, depending on the scale and the directional partition. The best results were obtained with neighborhoods that include only a parent and the eight nearest neighbors, and this choice will be referred to hereunder.

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 discuss first the Basis-Pursuit De-Noising (BPDN) method, which was introduced by Chen, Donoho and Saunders [5]. It refers to the solution of

$$\hat{\mathbf{z}} = \min_{\mathbf{z}} \frac{1}{2}\|y - \Phi\mathbf{z}\|_2^2 + \lambda\|\mathbf{z}\|_1, \quad (2)$$

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, 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

$$\hat{\mathbf{z}} = \min_{\mathbf{z}} \frac{1}{2}\|y - \Phi\mathbf{z}\|_2^2 + \sum_i \lambda_i|z_i|. \quad (3)$$

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.

Sendur and Selesnick [4] 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. We can easily extend their model 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

$$p(\mathbf{z}) = K\exp\left(-a\sqrt{\sum_j\left(\frac{z_j}{\sigma_j}\right)^2}\right), \quad (4)$$

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. (4), it can be compared with an empirical histogram. Figure 2 shows (in white) the log joint histogram of a reference coefficient and one of its nearest neighbors, estimated from the finest CT bands of several images. 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. This can be solved by adding a small positive constant ε into the square root of Eq. (4). 2) The decay rate diminishes as $|\mathbf{z}|$ increases, while the model suggests a constant decay rate. Rectifying this difference is obtained by decreasing the power inside the exponent from $1/2$ to $1/\gamma \quad (\gamma > 2)$. Thus, the modified model is given by

$$p(\mathbf{z}) = K \exp\left(-a\left(\sum_j \left(\frac{z_j}{\sigma_j} \right)^2 + \varepsilon \right)^{\frac{1}{\gamma}} \right). \quad (5)$$

Manual fitting between the estimated and the modelled log joint pdf resulted in $\varepsilon = 2.5 \cdot 10^{-2}, \gamma = 6$ (see Fig. 2).

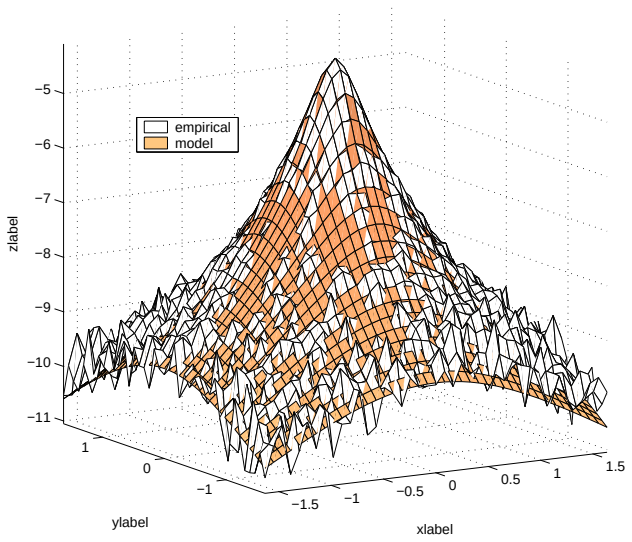

Fig. 2. Log joint histogram of two nearest neighbors: empirical vs. the proposed model in (5). The axes are xlabel=z_1 (reference), ylabel=z_2 (neighbor), and zlabel=$lnp(z_1, z_2)$.

The question arising now is how to extend the local prior model into a global one. Sendur and Selesnick [4] assumed independent neighborhoods, to simplify mathematical manipulations. In a similar fashion, we also embrace the independency assumption. Incorporating this supposition into Eq. (2), we get

$$\hat{\mathbf{z}} = \min_z \ \frac{1}{2}\|y - \Phi\mathbf{z}\|_2^2 + \lambda \sum_i \left(\sum_{j(i)} \left(\frac{z_j}{\sigma_j} \right)^2 + \varepsilon \right)^{\frac{1}{\gamma}}, \quad (6)$$

where λ is again an adjustable constant. Here we denote $\mathbf{z}$ as the global coefficients vector and $\{j(i)\}$ as the indices of the coefficients included in the i-th neighborhood. Note that the outer summation in Eq. (6) is not made over the lowpass coefficients, since the discussed dependency model is not valid for these. This method will be denoted hereafter BPDN-VAR.

4.2. Variance Estimation

The implementation of the new algorithm requires an estimation of the variances $\{\sigma_i\}$ from the given data. One common way of estimating the variances is by using coefficients from the reference coefficient's vicinity. 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.* [11] 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.

$$\mathbf{w}_{LS} = (Z^t Z)^{-1} Z^t |\mathbf{z}|, \quad (7)$$

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.

Next, the contexts $\{y_j\}$ in each subband 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 given by

$$\hat{\sigma}_i^2 = \max\left\{ \frac{1}{2L+1} \sum_{j(i)} z_j^2 - \sigma_{n,i}^2 \quad , 0 \right\}, \quad (8)$$

where $\sigma_{n,i}^2$ is the noise variance at the i-th coefficient (it is in fact constant over a subband). As Fig. 1 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. [11], we choose $L = \max\{100, 0.02M\}$ to guarantee reliable estimation along with adaptivity to varying characteristics, as well as $p = 9$ (eight spatial neighbors and one parent).

5. EXPERIMENTS

5.1. Implementation Issues

The BPDN-VAR method is specified by 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. 2). However, such a choice might not be suitable for higher dimensional distributions. Moreover, for $\gamma > 2$ the objective function in Eq. (6) 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, the choice which led to the best performance was of a parent and the four nearest spatial neighbors. 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). In addition, to show the effect of spatial adaptivity, a special case of BPDN-VAR-NN (denoted by BPDN-BAND), where the parameters $\{\sigma_i\}$ are only subband-dependent, will be tested.

Returning briefly to Eq. (3), 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. (6)), each coefficient appears either five, six or nine times in the summation, depending if it belongs to the finest scale, the second-finest scale, or any other scale, respectively. 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 $(9/5)^2$ or $(6/5)^2$ (for the coarsest scales and the second-finest scale, respectively), and also to set $\lambda = \lambda_0/5$. We should note that the PSNR values remain virtually unchanged for $\lambda \in [\lambda_0/6, \lambda_0/3]$, yet better visual quality was obtained with $\lambda_0/3$, which was thus chosen.

Following the selection and estimation of the unknown parameters, the minimization of the cost function in Eq. (6) can begin. A work by Elad [1] showed that the BPDN problem (Eq. (2)) 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. 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.

After testing many optimization algorithms, we finally decided to use the Truncated-Newton algorithm with preconditioning [12] for BPDN-VAR method. More details about the optimization method, including explicit expressions for the gradient and the hessian's diagonal, can be found in Ref. [13]. In our simulations the PSNR increased with every multidimensional iteration, until it settled after about 20 iterations.

For further comparison, we have also tested 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. [7], we set $K = 4$ for the finest scale, and $K = 3$ otherwise. For all of the discussed methods, we used a directional partition of $8, 8, 16$ and 16 directions (from coarse to fine).

5.2. Results

Figure 3 displays a comparison between the BPDN-VAR and BLS-GSM methods, for a 200×200 slice of *Peppers*. The corresponding PSNR values appear in Table 1. Although the PSNR of BLS-GSM is slightly higher, the visual quality of both methods is similar.

Table 1 summarizes the PSNR results of all of the examined methods, for $\sigma_n = 20$. This comparison reveals some interesting observations: 1) Spatial adaptivity improves the performance dramatically, as the comparison between BPDN-BAND and BPDN-VAR-NN shows. 2) The BPDN-VAR method surpasses BPDN-VAR-NN uniformly ($0.38dB$ on average). Thus, modifying the prior to account for the dependencies is worthwhile. 3) BPDN-VAR method attains roughly the same PSNR as BLS-GSM (merely $0.06dB$ less).

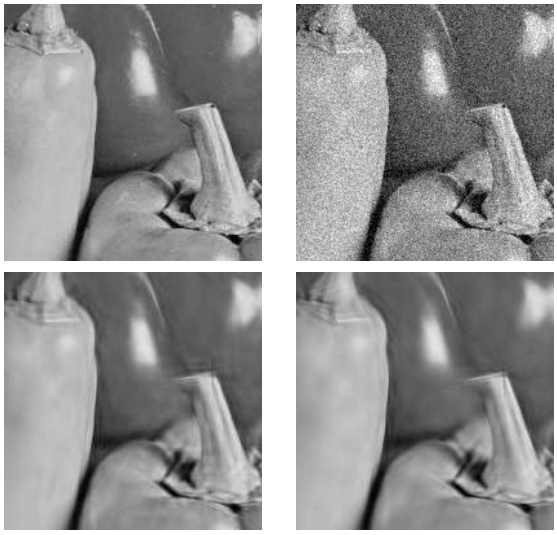

Fig. 3. Denoising results of a 200×200 slice of *Peppers* (for $\sigma_n = 20$). From left to right and top to bottom: Original; Noisy; BLS-GSM; BPDN-VAR.

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

	Peppers256	*Peppers*	*Lena*	*Barbara*
HT	28.21	30.87	31.46	28.36
BPDN-BAND	27.96	30.00	30.30	27.58
BPDN-VAR-NN	29.21	31.14	31.49	29.61
BPDN-VAR	**29.43**	31.57	32.02	29.94
BLS-GSM	29.27	**31.69**	**32.06**	**30.19**

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 superior to the classic Basis-Pursuit Denoising (BPDN), which does not account for these dependencies. Even though the new prior still does not describe accurately the true probability function (because of the neighborhoods independency assumption), it does provide a step forward in that direction.

As for the future work plan, we mention several topics: 1) testing various values of γ and ε (see Sect. 4.1). 2) Extension to more general inverse problems such as deblurring. 3) Further modification of the prior in Eq. (6), in order to better describe the inter-coefficient dependencies.

7. ACKNOWLEDGEMENTS

The authors thank Mr. Yue Lu for supplying his Contourlet-SD implementation.

8. REFERENCES

[1] Micael Elad, "Why simple shrinkage is still relevant for redundant representations?," *IEEE Trans. on Information Theory*, submitted, Jan. 2005.

[2] Duncan D. Y. Po and Minh N. Do, "Directional multiscale modeling of images using the contourlet transform," *IEEE Trans. on Image Processing*, to appear, 2005.

[3] 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*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.

[4] Levent Sendur and Ivan W. Selesnick, "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency," *IEEE Trans. Signal Processing*, vol. 50, no. 11, pp. 2744–2755, Nov. 2002.

[5] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[6] Francoise Dibos and Georges Koepfler, "Global total variation minimization," *SIAM Journal on Numerical Analysis*, vol. 37, no. 2, pp. 646–664, 2000.

[7] Minh N. Do and Martin Vetterli, "The contourlet transform: An efficient directional multiresolution image representation," *IEEE Trans. on Image Processing*, to appear, 2004.

[8] Emmanuel J. Candés and David Donoho, "New tight frames of curvelets and optimal representations of objects with smooth singularities," Tech. Rep., 2002.

[9] Boaz Matalon, Michael Zibulevsky, and Michael Elad, "Image denoising with the contourlet transform," in *Proceedings of the SPIE conference wavelets*, July 2005, vol. 5914.

[10] Yue Lu and Minh N. do, "Constructing contourlets with spatial/frequency localization," to appear, 2005.

[11] S. Grace Chang Bin Yu and Martin Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," *IEEE Trans. on Image Processing*, vol. 9, no. 9, pp. 1522–1531, Sep. 2000.

[12] Stephen G. Nash, "A survey of truncated-newton methods," *Journal of Computational and Applied Mathematics*, vol. 124, pp. 45–59, 2000.

[13] Boaz Matalon, Michael Zibulevsky, and Michael Elad, "A new method for image denoising," Tech. Rep., to appear, 2006.