A Discriminative Approach for Wavelet Denoising

Yacov Hel-Or and Doron Shaked

***Abstract*—This paper suggests a discriminative approach for wavelet denoising where a set of mapping functions (MFs) are applied to the transform coefficients in an attempt to produce a noise free image. As opposed to the descriptive approaches, modeling image or noise priors is not required here and the MFs are learned directly from an ensemble of example images using least-squares fitting. The suggested scheme generates a novel set of MFs that are essentially different from the traditional soft/hard thresholding in the over-complete case. These MFs are demonstrated to obtain comparable performance to the state-of-the-art denoising approaches. Additionally, this framework enables a seamless customization of the shrinkage operation to a new set of restoration problems that were not addressed previously with shrinkage techniques, such as deblurring, JPEG artifact removal, and various types of additive noise that are not necessarily Gaussian white noise.**

***Index Terms*—Image deblurring, image denoising, JPEG artifact removal, shrinkage, wavelet.**

I. Introduction

MANY imaging devices that acquire or process digital images introduce artifacts in the processing pipeline. These artifacts include additive noise, image blurring, compression artifacts, missing pixels, geometric distortions, etc. Image restoration is an attempt to reduce such artifacts using postprocessing operations. One important topic in image restoration deals with image denoising, where noisy observations of images are attempted to be cleaned. In this paper, we focus on denoising images contaminated with additive noise whose statistical distribution is known. Consider a noisy image

$$\mathbf{y} = \mathbf{x} + \mathbf{n} \tag{1}$$

where $\mathbf{y}$ is the observed image, $\mathbf{x}$ the unknown original image and $\mathbf{n}$ the contaminating noise (all in vector notation). The goal is to reconstruct the original image $\mathbf{x}$ given the noisy measurement $\mathbf{y}$. This problem is a typical instance of an inverse problem where the solution must consider prior knowledge of the distribution of $\mathbf{x}$. Hence, the prior distribution of natural images or of any other specific class of images plays a key role in any denoising approach.

A common approach for modeling the statistical prior of natural images is to estimate their statistical distribution in a transform domain. This is often implemented using some type of wavelet transform. The main motivation for this approach stems from the observation that the wavelet transform of natural images tends to reduce coefficient dependencies [1]–[4]. Hence, it is possible to make a reasonable inference about the joint distribution of the wavelet coefficients from their marginal distributions. When dealing with image denoising, this leads to a family of classical techniques known as the *wavelet shrinkage methods* introduced by Donoho and Johnstone in 1994 [5]–[7]. These techniques amount to modifying the coefficients in the transform domain using a set of scalar mapping functions, $\{\mathcal{M}^i\}$, called *mapping functions* (MFs). The MFs are also known as *shrinkage functions* since they commonly apply an adaptive shrinking operation to the transform coefficients. The shrinkage approach is comprised of a wavelet transform

$$\mathbf{y}_{\mathrm{W}} = \mathbf{W}\mathbf{y} \tag{2}$$

followed by a correction step in which the wavelet coefficients are rectified according to a set of MFs

$$\hat{\mathbf{x}}_{\mathrm{W}} = \vec{\mathcal{M}}_{\mathrm{W}}\{\mathbf{y}_{\mathrm{W}}\} \tag{3}$$

where $\vec{\mathcal{M}}_{\mathrm{W}} = [\mathcal{M}^1_{\mathrm{W}}, \mathcal{M}^2_{\mathrm{W}}, \ldots]$ is a vector of scalar mapping functions. The denoised image is obtained by applying the inverse transform to the modified coefficients

$$\hat{\mathbf{x}} = \mathbf{W}^{-1}\hat{\mathbf{x}}_{\mathrm{W}}. \tag{4}$$

Due to their simplicity and good results, shrinkage approaches have received a great deal of attention over the last decade. Hundreds of shrinkage methods have been proposed differing mostly in the type of transform used and in the form in which the MFs are applied. The justification for applying a marginal (scalar) MF to each coefficient independently can be shown to emerge from the independence assumption of the wavelet coefficients where the noise is modeled as white and Gaussian. This assumption was postulated in the early studies in which MFs were applied to unitary transforms.

Since the pioneering work of Donoho and Johnston, various efforts have been made to improve the denoising results of wavelet based methods. Such efforts generally concentrated on two main directions. The first direction attempts to improve the results by abandoning the unitary representation and working in over-complete transform domains. Such transforms include undecimated wavelets [8], steerable wavelets [9], and other recently developed transforms, such as ridgelets [10]–[12], contourlets [13], [14], curvelets [15], and image dependent transforms [16]. These transforms were shown to better represent natural images in the sense that their coefficients exhibit better sparsity. Additionally, the over-completeness was shown to significantly improve denoising results in addition to having shift-invariant performance. Although the independence assumption can no longer be justified in the over-complete

Manuscript received January 15, 2007; revised August 22, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Pier Luigi Dragotti.

Y. Hel-Or is with the Interdisciplinary Center, Efi Arazi School of Computer Science, Herzliya, Israel 46150 (e-mail: toky@idc.ac.il).

D. Shaked is with the Hewlett-Packard Labs Israel, Technion City, Haifa, Israel (e-mail: doron.shaked@hp.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2008.917204

domain, most of the conventional methods naively borrowed the traditional MFs from the unitary case.

The second direction towards improvement relaxed the independence assumption of the signal's wavelet coefficients and concentrated on modeling the statistical dependencies between neighboring coefficients. This scheme can be seen as diverging from the scalar MFs to multivariate MFs where transform coefficients (an individual or a group) are rectified according to a group of measured coefficients. Intercoefficient dependencies are exploited using any of a range of techniques, such as the joint sparsity assumption [17], [18], HMM and Bayesian models [19]–[23], context modeling [24], [25], tree models representing parent-child dependencies [26], co-occurrence matrix [27], adaptive thresholding [28], [29], geometrical prior models [30], and more. These types of techniques achieve very good denoising performance; however, they generally lack the efficiency and simplicity of the classical shrinkage approaches.

Common to all the conventional techniques for generating MFs, regardless of the approach used, is that the MFs are derived in a *descriptive* manner. Namely, a statistical model is first constructed describing the statistical prior of the transform coefficients. Based on this prior, a set of MFs are derived (scalar or multivariate, parametric or nonparametric) which are designed to rectify the contaminated coefficients. Clearly, imprecise modeling of the statistical prior directly leads to a deterioration in the resulting performance. Because intercoefficient dependencies are too complicated to model, in particular, in the over-complete case, it is expected that the statistical models are far from precise. And indeed, due to the high dimensionality of the joint probability, ad-hoc assumptions have been commonly made in order to make the problem tractable. Such assumptions include, e.g., ignoring the intercoefficient dependencies (e.g., [5] and [9]), modeling only bivariate or parent-child dependencies (e.g., [26]), and modeling the joint dependencies of a small group of neighboring coefficients but assuming simplified parametric models (e.g., [25]).

This paper suggests a new scheme for designing a set of MFs using a *discriminative* framework. In contrast to the conventional approaches, this technique does not require any estimation of the prior model nor the noise characteristics. Rather, a set of MFs is constructed using an ensemble of example images whose clean and contaminated versions are supplied offline. The MFs are designed to perform "optimally" with respect to the given examples, under the assumption that they will perform equally well with similar new examples.

The suggested approach retains the traditional scalar MFs that are applied to each wavelet coefficient independently. Nevertheless, although the MFs are applied in a marginal manner, their construction is affected by intercoefficient dependencies. In fact, it is shown that the obtained MFs differ essentially from the conventional monotonic hard/soft thresholding functions. Moreover, despite the fact that scalar MFs are used, the denoising results are comparable and sometimes even better than the state-of-the-art multivariate prior based techniques. Thus, the suggested approach, while maintaining the simplicity and efficiency of the scalar shrinkage approaches, typically does not compromise the resulting quality.

The advantages of the proposed scheme stem, in part, from the following characteristics.

- First, the MFs are constructed in an optimal manner taking into account intercoefficient dependencies. Although the MFs apply nonlinear operations, their construction is performed in a closed form solution using a spline based representation.
- The second source of improvement stems from the optimality criteria applied in this method. While most shrinkage approaches construct the MFs using the MAP criterion that rely on imprecise prior models, the proposed method uses a least-squares (LS) scheme approximating the MFs directly from an example set. Thus, the suggested scheme avoids the need for modeling complex statistical prior in high-dimensional space.
- The third source of improvement is due to the domain in which the optimality criterion is preferably performed. In the suggested method, the objective goal is expressed in the spatial domain, which is the domain within which images are perceived. Most wavelet shrinkage approaches use optimality criteria expressed in the wavelet domain. While a transform-domain optimization criterion is justified in unitary transforms, it does not properly extend to over-complete transforms. Furthermore, it can be shown that the optimal solution in the over-complete transform domain does not guarantee optimality in the spatial domain (see Appendix in [31]).

The approach suggested in this paper is presented in the context of denoising. However, using the proposed discriminative framework, other reconstruction problems that were not previously addressed with shrinkage approaches, can be dealt with seamlessly, as long as the reconstruction process involves scalar look-up-tables applied in the wavelet domain. These include, e.g., image deblurring, JPEG artifact removal, and various types of additive noise. Some results will be shown for these cases.

The rest of the paper is organized as follows. The next two sections describe the classical shrinkage approaches in the unitary and the over-complete domains. These sections provide the background for our proposed method. In Section IV, the slice transform (SLT) is introduced along with its properties. Section V presents the proposed method, and Section VI addresses several computational issues. Simulation results as well as implementations in other restoration problems are presented in Section VII.

II. Restoration in Unitary Transform Domains

The justification for using scalar mapping functions can be shown to emerge from the MAP estimation and the independence assumption of the wavelet coefficients. Consider a degradation model as described in (1). The MAP solution $\hat{\mathbf{x}}(\mathbf{y})$ is the image that maximizes the *a posteriori* probability

$$\hat{\mathbf{x}}(\mathbf{y}) = \arg\max_{\mathbf{x}} P(\mathbf{x} \,|\, \mathbf{y})$$

This maximization can be expressed in the wavelet domain, as well. Denoting a *unitary* wavelet transform $\mathbf{x}_{\mathrm{W}} = W\mathbf{x}$ and $\mathbf{y}_{\mathrm{W}} = W\mathbf{y}$, the MAP estimation gives

$$\hat{\mathbf{x}}_{\mathrm{W}}(\mathbf{y}_{\mathrm{W}}) = \arg\max_{\mathbf{x}_{\mathrm{W}}} P(\mathbf{x}_{\mathrm{W}} \,|\, \mathbf{y}_{\mathrm{W}}). \tag{5}$$

Using the Bayes conditional rule and exploiting the monotonicity of the log function, the maximization in (5) is equivalent to

$$\begin{aligned}\hat{\mathbf{x}}_{\mathrm{W}} &= \arg\max_{\mathbf{x}_{\mathrm{W}}} P(\mathbf{y}_{\mathrm{W}} \mid \mathbf{x}_{\mathrm{W}}) P(\mathbf{x}_{\mathrm{W}}) \\ &= \arg\min_{\mathbf{x}_{\mathrm{W}}} \{-\log P(\mathbf{y}_{\mathrm{W}} \mid \mathbf{x}_{\mathrm{W}}) - \log P(\mathbf{x}_{\mathrm{W}})\}. \quad (6)\end{aligned}$$

The first term, $\log P(\mathbf{y}_{\mathrm{W}} \mid \mathbf{x}_{\mathrm{W}})$, is referred to as the *likelihood term*. It depends solely on the noise characteristics. In the case of white Gaussian noise, this term reduces to

$$\begin{aligned}-\log P(\mathbf{y}_{\mathrm{W}} | \mathbf{x}_{\mathrm{W}}) &= \eta \|\mathbf{x}_{\mathrm{W}} - \mathbf{y}_{\mathrm{W}}\|^2 \\ &= \eta \sum_i \left\|x_{\mathrm{W}}^i - y_{\mathrm{W}}^i\right\|^2 \quad (7)\end{aligned}$$

where x_{W}^i and y_{W}^i denote the ith elements of the corresponding vectors and η is a constant depending on the noise variance. The second term in (6), $\log P(\mathbf{x}_{\mathrm{W}})$, is known as the *regularization term* or the *prior term* as it specifies the *a priori* probability of the original image $\mathbf{x}_{\mathrm{W}}$. Taking into account the independence assumption of the wavelet coefficients, the second term can be rewritten as

$$\begin{aligned}-\log P(\mathbf{x}_{\mathrm{W}}) &= -\log \prod_i P_i(x_{\mathrm{W}}^i) \\ &= -\sum_i \log P_i\left(x_{\mathrm{W}}^i\right). \quad (8)\end{aligned}$$

Substituting (7) and (8) into (6), the overall minimization amounts to a set of independent scalar minimizations, each of which corresponds to a particular coefficient

$$\hat{x}_{\mathrm{W}}^i\left(y_{\mathrm{W}}^i\right) = \arg\min_{x_W^i} \left\{\eta \left\|x_{\mathrm{W}}^i - y_{\mathrm{W}}^i\right\|^2 - \log P_i\left(x_{\mathrm{W}}^i\right)\right\} \quad \forall i. \quad (9)$$

The last expression gives the justification for applying a scalar MF to each wavelet coefficient independently. Each value y_{W}^i is mapped to: $\hat{x}_{\mathrm{W}}^i = \mathcal{M}_{\mathrm{W}}^i\{y_{\mathrm{W}}^i\}$ which is given in (9). Note that for a particular noise variance, the variations in the MFs, $\mathcal{M}_{\mathrm{W}}^i\{\cdot\}$, depend solely on $P_i(x_{\mathrm{W}}^i)$. Furthermore, assuming the statistics of natural images are homogeneous [32], it implies that all wavelet coefficients belonging to a particular wavelet subband share the same distribution. Namely, w.l.o.g. if a coefficient x_{W}^i belongs to the jth subband so that $j = \mathrm{band}(i)$, we have

$$P_i\left(x_{\mathrm{W}}^i\right) = P_{\mathrm{band}(i)}\left(x_{\mathrm{W}}^i\right)$$

and, consequently

$$\hat{x}_{\mathrm{W}}^i = \mathcal{M}_{\mathrm{W}}^{\mathrm{band}(i)}\left\{y_{\mathrm{W}}^i\right\}. \quad (10)$$

Thus, if the wavelet transform is composed of K subbands, only K distinct MFs must be evaluated. To emphasize this point and simplify notations in the rest of the paper, we reorder the rows of the Wavelet transform W in (2) so that transform rows corresponding to a wavelet subband are co-located in a block. Naturally, we extend the same reordering to $\mathbf{y}_{\mathrm{W}}$. Assuming a corresponding permutation matrix P

$$B = PW = \begin{bmatrix} B_1 \\ \vdots \\ B_K \end{bmatrix} \quad \text{and} \quad B\mathbf{y} = \mathbf{y}_{\mathrm{B}} = \begin{bmatrix} \mathbf{y}_{\mathrm{B}_1} \\ \vdots \\ \mathbf{y}_{\mathrm{B}_K} \end{bmatrix} \quad (11)$$

where $\mathbf{y}_{\mathrm{B}_k}$ represents the coefficients in the kth subband. In the new reordering, a vector of MFs, $\vec{\mathcal{M}}_{\mathrm{B}} = [\mathcal{M}_{\mathrm{B}}^1, \mathcal{M}_{\mathrm{B}}^2, \ldots, \mathcal{M}_{\mathrm{B}}^K]$, is applied as follows. Since $\mathcal{M}_{\mathrm{B}}^k$ is applied individually to all coefficients in the kth subband, the estimated image of (4) is rewritten as

$$\hat{\mathbf{x}} = B^T \vec{\mathcal{M}}_{\mathrm{B}}\{\mathbf{y}_{\mathrm{B}}\} = \sum_{k=1}^{K} B_k^T \mathcal{M}_{\mathrm{B}}^k\{\mathbf{y}_{\mathrm{B}_k}\}. \quad (12)$$

The main scope of this paper is the optimal estimation of the MFs $\mathcal{M}_{\mathrm{B}}^k\{\cdot\}$. There is a wealth of papers dealing with the estimation of the MFs in the context of denoising. The early studies of Donoho and Johnston suggested using *soft thresholding* or *hard thresholding* as shrinkage functions [6], [7]. These can be shown to emerge from the MAP estimation where the distributions of the wavelet coefficients are generalized Gaussians (GGD): $P(x) \sim e^{-|x/s|^p}$. Soft-thresholding is a result of assuming a Laplacian distribution (i.e., $p = 1$) while hard-thresholding assumes a sharper distribution with $p = 0.5$ [20], [33]. Later studies extended the thresholding approach to other distributions adapting the MFs to these distributions (e.g., [9], [20], [33]–[35]).

III. Restoration in Over-Complete Domains

Although the shrinkage approach using unitary transforms provides good results, significant improvement is achieved when implementing this technique in over-complete representations. In most cases, this is implemented using the un-decimated wavelet transform or any other shift-invariant transforms (sliding local DCT, ridgelets, contourlets, steerable wavelet, etc.). Adopting the notation of (11) for the over-complete case, the image transform is given by $\mathbf{y}_{\mathrm{B}} = B\mathbf{y}$, where B is composed of distinct subband matrices. Unlike the unitary transform, however, the number of rows in B is larger than the dimensionality of the signal $\mathbf{y}$. Modifying $\mathbf{y}_{\mathrm{B}}$ using a vector of MFs $\vec{\mathcal{M}}_{\mathrm{B}}\{\mathbf{y}_{\mathrm{B}}\}$ aims at removing the noise components. Hence, it is assumed that

$$B\hat{\mathbf{x}} = \vec{\mathcal{M}}_{\mathrm{B}}\{B\mathbf{y}\}$$

and the image is reconstructed using the pseudo-inverse

$$\begin{aligned}\hat{\mathbf{x}} &= (B^T B)^{-1} B^T \vec{\mathcal{M}}_{\mathrm{B}}\{\mathbf{y}_{\mathrm{B}}\} \\ &= (B^T B)^{-1} \sum_{k=1}^{K} B_k^T \mathcal{M}_{\mathrm{B}}^k\{\mathbf{y}_{\mathrm{B}_k}\} \quad (13)\end{aligned}$$

where we have

$$(B^T B)^{-1} = \left(\sum_{k=1}^{K} B_k^T B_k\right)^{-1}.$$

If the transform B is tight frame, (13) can be simplified due to the fact that $(1/N)B^TB = I$.

Viewing the shrinkage approach as a set of mapping functions applied to each subband independently may suggest that the MFs applied in the over-complete case are correspondingly similar to those applied in a unitary case (9). And indeed this attitude was broadly adopted in previous works [8], [11], [14], [15], [36]. However, even if we may assume statistical independence in the coefficients belonging to a unitary transform, it definitely cannot be extended to coefficients in the over-complete transform where the transform coefficients are inherently dependent. Thus, a new set of MFs must be designed that takes into consideration the interband dependencies.

Another issue that causes the over-complete case to differ from the unitary case is the domain in which the minimization criterion is applied. To clarify this point, consider finding the optimal MFs for the *unitary* case with respect to the MSE criterion. Namely, finding $\vec{\mathcal{M}}_{\mathrm{B}}$ that minimizes

$$\varepsilon = E_{\mathbf{x}|\mathbf{y}}\{\|\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}\|^2\}$$

where we have determined that $\hat{\mathbf{x}}(\mathbf{y}) = B^T\vec{\mathcal{M}}_{\mathrm{B}}\{B\mathbf{y}\}$, and $E_{\mathbf{x}|\mathbf{y}}\{\cdot\}$ stands for the conditional expectation of $\mathbf{x}$ given $\mathbf{y}$. Whenever B is unitary, this minimization can be expressed equivalently in the transform domain, namely

$$E_{\mathbf{x}|\mathbf{y}}\{\|B^T\vec{\mathcal{M}}_{\mathrm{B}}\{B\mathbf{y}\} - \mathbf{x}\|^2\} = E_{\mathbf{x}|\mathbf{y}}\{\|\vec{\mathcal{M}}_{\mathrm{B}}\{B\mathbf{y}\} - B\mathbf{x}\|^2\}.$$

However, for an over-complete transform this equality is not valid anymore (see [31] for more details), which implies that the optimization for $\vec{\mathcal{M}}_{\mathrm{B}}$ should be expressed in the spatial domain. Due to the fact that the inverse transform couples wavelet coefficients, spatial domain optimization is far more complex.

Although scalar MFs may no longer be justified when the transform coefficients are mutually dependent, the superior results of applying scalar MFs in the over-complete case suggest that such a scheme is still very useful in addition to its appealing efficiency. Furthermore, in a recent paper, Elad [36] justifies scalar MFs as being the first step in an iterative minimization scheme. Justified as the optimal solution or not, there is definitely an interest in finding the best MFs in the over-complete case while considering intercoefficient dependencies. To the best of our knowledge, the optimal design of MFs in the over-complete domain was not discussed in the literature, and in most cases the applied MFs were naively borrowed from the unitary case.

This paper, presents a new scheme for image denoising in the over-complete case, where MFs are represented in a linear manner using a spline representation which we call a *slice transform*. As an introduction to the proposed approach we first introduce the SLT of an image. This representation will be used in later sections to calculate the optimal MFs.

IV. Slice Transform and Its Properties

Let $x \in [a, b) \in \mathcal{R}$ be a real value in the half open interval $[a, b)$. The interval is divided into M bins whose boundaries form a vector $\mathbf{q}$

$$\mathbf{q} = [q_0, q_1, \ldots, q_M]^T$$

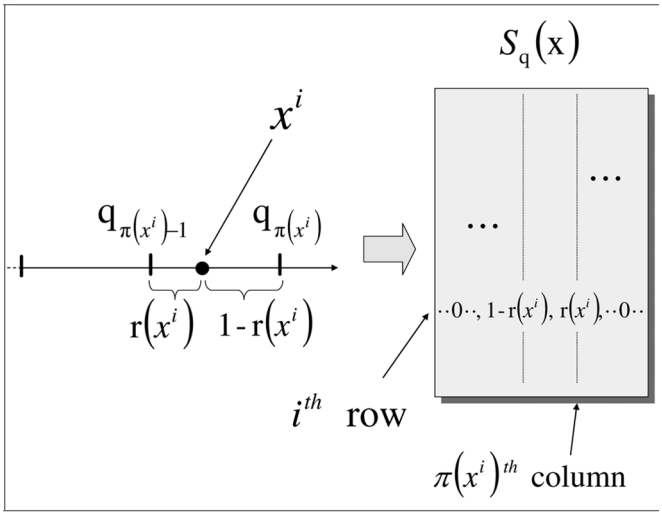

Fig. 1. Representing x as a quantized value and residue.

such that

$$q_0 = a < q_1 < q_2 \cdots < q_M = b.$$

The value x is naturally associated with a single bin $\pi(x) \in \{1 \cdots M\}$ and a corresponding normalized residue, $r(x)$, where (see Fig. 1)

$$\pi(x) = j \text{ if } x \in [q_{j-1}, q_j)$$

and

$$r(x) = \frac{x - q_{\pi(x)-1}}{q_{\pi(x)} - q_{\pi(x)-1}}.$$

Note that $r(x) \in [0, 1)$, where $r(x) = 0$ if $x = q_{\pi(x)-1}$, and $r(x) \to 1$ if $x \to q_{\pi(x)}$. The value x can then be expressed as a linear combination of $q_{\pi(x)}$ and $q_{\pi(x)-1}$

$$x = r(x)q_{\pi(x)} + (1 - r(x))q_{\pi(x)-1}. \tag{14}$$

Equation (14) can be rewritten in vectorial form

$$x = S_{\mathbf{q}}(x)\mathbf{q} \tag{15}$$

where $S_{\mathbf{q}}(x)$ is defined as an $M + 1$ dimensional row vector as follows:

$$S_{\mathbf{q}}(x) = [0, \ldots, 0, 1 - r(x), r(x), 0, \ldots, 0]$$

and where the values $1-r(x)$ and $r(x)$ are located in the $(\pi(x)-1)^{\mathrm{th}}$ and $\pi(x)^{\mathrm{th}}$ entries, respectively. We now define a vectorial extension of (15). Let $\mathbf{x}$ be an N- dimensional vector whose elements satisfy $x^i \in [a, b)$. The SLT of $\mathbf{x}$ is defined as follows:

$$\mathbf{x} = S_{\mathbf{q}}(\mathbf{x})\mathbf{q} \tag{16}$$

where $S_{\mathbf{q}}(\mathbf{x})$ is an $N \times (M + 1)$ matrix defined as follows (see Fig. 1)

$$[S_{\mathbf{q}}(\mathbf{x})](i, j) = \begin{cases} r(x^i), & \text{if } \pi(x^i) = j \\ 1 - r(x^i), & \text{if } \pi(x^i) = j + 1 \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

The matrix $S_{\mathbf{q}}(\mathbf{x})$ has N rows, each corresponding to an entry in $\mathbf{x}$, and $M + 1$ columns associated with the bin boundaries

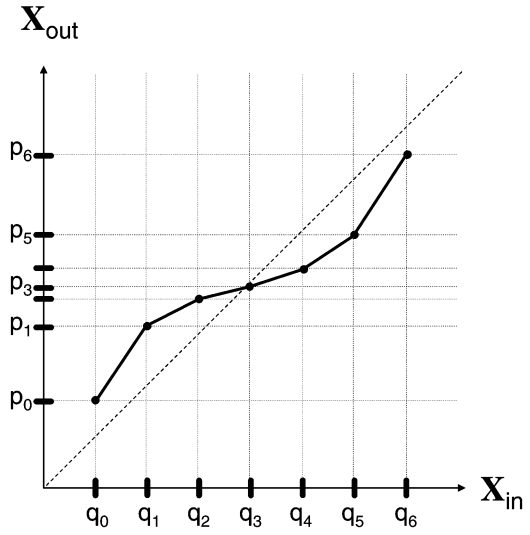

Fig. 2. Illustration of a piecewise linear map in accordance with the substitution property of the SLT.

defined in $\mathbf{q}$. The SLT (16) is eventually a linear representation of a signal $\mathbf{x}$ where the basis functions are linear splines. A unique property of the SLT is the *substitution property*.

Proposition: Substituting the boundary vector $\mathbf{q}$ with a different vector $\mathbf{p}$ performs a piecewise linear mapping of the values in $\mathbf{x}$

$$\mathcal{M}_{\mathbf{q},\mathbf{p}}\{\mathbf{x}\} = S_{\mathbf{q}}(\mathbf{x})\mathbf{p}$$

where $\mathcal{M}_{\mathbf{q},\mathbf{p}}\{\mathbf{x}\}$ is such that values $\{x \in [q_{j-1}, q_j)\}$ are mapped linearly to the interval $[p_{j-1}, p_j)$. This means that for every $\alpha \in [0,1)$, and $j \in \{1,2,\dots M\}$ the value $x = \alpha q_j + (1-\alpha)q_{j-1}$ is mapped to $\mathcal{M}_{\mathbf{q},\mathbf{p}}\{x\} = \alpha p_j + (1-\alpha)p_{j-1}$ (see Fig. 2 for an illustration of such a mapping).

The substitution property is the key principle behind the approach suggested in this paper. Namely, expressing a family of nonlinear MFs in a linear matrix form. This, in turn, enables a simple optimization of the MFs as a solution to a linear set of equations. Thus, if we are willing to approximate general nonlinear maps as piece-wise linear maps, we can obtain the optimal (piece-wise linear) map. Note that one may always use a finer quantization grid that will result in a better approximation of the desired optimal map.

V. Estimating the Mapping Functions

Consider the restoration scheme in the over-compete domain and recall that our main goal is to find a vector of MFs $\vec{\mathcal{M}}_{\mathrm{B}} = [\mathcal{M}_{\mathrm{B}}^1, \mathcal{M}_{\mathrm{B}}^2, \dots]$ that would best restore $\mathbf{x}$ from $\mathbf{y}_{\mathrm{B}}$ using (13)

$$\hat{\mathbf{x}}(\mathbf{y}) = (B^T B)^{-1} \sum_{k=1}^{K} B_k^T \mathcal{M}_{\mathrm{B}}^k \{\mathbf{y}_{\mathrm{B}_k}\}. \tag{18}$$

If we are willing to restrict our MFs to be a piecewise linear map, we may apply the substitution property of the SLT and obtain

$$\mathcal{M}_{\mathrm{B}}^k \{\mathbf{y}_{\mathrm{B}_k}\} \approx \mathcal{M}_{\mathbf{q}_k,\mathbf{p}_k}\{\mathbf{y}_{\mathrm{B}_k}\} = S_{\mathbf{q}_k}(\mathbf{y}_{\mathrm{B}_k})\,\mathbf{p}_k$$

where $\mathcal{M}_{\mathbf{q}_k,\mathbf{p}_k}$ describes the piecewise linear approximation of the mapping $\mathcal{M}_{\mathrm{B}}^k$. Using (18), the resulting image is then given by

$$\hat{\mathbf{x}}(\mathbf{y}) = (B^T B)^{-1} \sum_{k=1}^{K} B_k^T S_{\mathbf{q}_k}(\mathbf{y}_{\mathrm{B}_k})\,\mathbf{p}_k = L(\mathbf{y}_{\mathrm{B}})\mathbf{p} \tag{19}$$

where now $\mathbf{p}$ is a $K(M+1)$ dimensional column vector constructed by stacking all $\mathbf{p}_k$ vectors together and $L(\mathbf{y}_{\mathrm{B}})$ is composed of all the SLT matrices

$$L(\mathbf{y}_{\mathrm{B}}) = (B^T B)^{-1} [H_1, H_2, \cdots, H_K] \tag{20}$$

where we define

$$H_i = B_i^T S_{\mathbf{q}_i}(\mathbf{y}_{\mathrm{B}_i}).$$

Note that in the undecimated transform cases, the term $B_k^T S_{\mathbf{q}_k}(\mathbf{y}_{\mathrm{B}_k})$ can be calculated efficiently by applying a 2-D convolution to each of the images composing the columns of $S_{\mathbf{q}_k}(\mathbf{y}_{\mathrm{B}_k})$.

The offline step of the proposed scheme aims at learning the optimal MFs to be applied. Namely, the goal is to find, for each k, the optimal $\mathbf{p}_k$ vector that together with the $\mathbf{q}_k$ vector defines the piecewise mapping functions $\mathcal{M}_{\mathbf{q}_k,\mathbf{p}_k}$. In the proposed scheme, the MFs are trained from an example set of clean signals $\{\mathbf{x}^e\}$ that are given along with their noisy counterparts $\{\mathbf{y}^e\}$. For simplicity, we first assume that a single signal $\mathbf{x}^e$ is given as an example along with its noisy version $\mathbf{y}^e$. The optimal (piecewise) MFs are obtained using a curve-fitting approach minimizing a LS criterion

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} \|\hat{\mathbf{x}}(\mathbf{y}^e) - \mathbf{x}^e\|^2. \tag{21}$$

Substituting (19) into the above equation gives rise to a closed-form solution

$$\hat{\mathbf{p}} = (L^T L)^{-1} L^T \mathbf{x}^e \quad \text{where} \quad L = L(\mathbf{y}_{\mathrm{B}}^e). \tag{22}$$

The above solution provides the optimal K MFs to be applied to the K wavelet bands respectively. In fact, the resulting MFs are designed to optimally reconstruct the clean example from its noisy counterpart. The obtained MFs are then used for denoising new signals that are assumed to have similar statistical characteristics as the training example (signal and noise). If possible, clean and noisy examples should be acquired prior and following to the degradation process (e.g., before and after a noisy channel, before and after JPEG compression). Another possibility is to model the degradation model and synthesize noisy examples from clean natural images. If both options are not available, it is always possible to approximate the noise variance using available methods for noise estimation (e.g., [37]–[39]) and synthesize noisy images based on the estimated noise characteristics. Note that this process is applied only once. Additionally, it is not required to train the MFs for each possible noise variance, since the MFs follow a scaling rule as will be elaborated in the next section.

VI. Implementation Considerations

The proposed optimal solution was detailed in (22). However, several computational issues are critical to the implementation of the approach. These will be addressed in this section.

Stabilizing the Solution: The first issue is related to the kurtotic distributions of the wavelet coefficients. In such distributions the vast majority of the coefficient values are close to zero while only a negligible fraction of the coefficients depart from zero. This behavior may give rise to over-fitting phenomena in the higher part of the mapping domain, where a small number of measured coefficients are available. In more severe cases there are quantization bins without any sample values at all, and the matrix $L^T L$ in (22) then becomes singular or ill-posed. In order to resolve this problem, one must incorporate a regularization term in the minimization scheme. Referring to (21) we add a regularization term as follows:

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} \{ \|\hat{\mathbf{x}}(\mathbf{y}^e) - \mathbf{x}^e\|^2 + \sqrt{\lambda}\|\mathbf{p} - \mathbf{q}\|^2 \}.$$

The regularization term biases the solution of $\mathbf{p}$ towards $\mathbf{q}$ (identity MFs), in particular, in those $\mathbf{p}$ entries which are associated with bins where limited or no measured data is available for them. This makes intuitive sense for large coefficients because it means they will have little shrinkage applied to them. The constant parameter λ controls the influence strength of the regularization term. It can be easily verified that the final solution of this system gives

$$\hat{\mathbf{p}} = (L^T L + \lambda I)^{-1}(L^T \mathbf{x}^e + \lambda \mathbf{q}) \tag{23}$$

where I denotes a $K(M+1) \times K(M+1)$ identity matrix. In order to maintain the influence of the measured data the regularization term should be kept as small as possible. In our implementation, we used $\lambda = (0.005N/M)^2$ where N is the number of image pixels, and N/M is the average number of pixels per quantization bin. Since we used a very weak regularization strength, this term influences only those quantization bins with very few or no sampling values at all. In those bins, the regularization term causes the respective coefficients to remain untouched. In fact, the regularization term stabilizes the shape of the MFs but its influence on the denoising performance is marginal.

Out-of-Range Coefficients: Another issue to address is how to deal with transform coefficients whose values fall outside the domain interval. Since the SLT transform assumes a limited range of transform coefficients, namely the range $[q_0, q_M)$, there might be cases where the coefficients fall outside this range. In such cases, we ignore the influence of these coefficients on the desired solution by adding a *residual term* to the SLT definition (16)

$$\mathbf{x} = S_{\mathbf{q}}(\mathbf{x})\mathbf{q} + \mathbf{h}$$

where the residual term $\mathbf{h}$ contains all entries in $\mathbf{x}$ whose values are outside the range $[q_0, q_M)$. In our restoration scheme, this gives

$$\mathcal{M}_{\mathbf{q}_k, \mathbf{p}_k}\left\{\mathbf{y}^e_{\mathrm{B}_k}\right\} = S_{\mathbf{q}_k}\left(\mathbf{y}^e_{\mathrm{B}_k}\right)\mathbf{p}_k + \mathbf{h}^e_k.$$

Inserting this term into (19) gives

$$\hat{\mathbf{x}}^e = L\mathbf{p} + \tilde{\mathbf{h}}^e \quad \text{where} \quad \tilde{\mathbf{h}}^e = (B^T B)^{-1} \sum_k B_k^T \mathbf{h}^e_k. \tag{24}$$

This also updates the final solution, which now reads

$$\hat{\mathbf{p}} = (L^T L + \lambda I)^{-1}(L^T(\mathbf{x}^e - \tilde{\mathbf{h}}^e) + \lambda \mathbf{q}). \tag{25}$$

Accordingly, during the restoration process, the piece-wise mapping functions are applied only to in-range coefficients while out-of-range coefficients are left untouched.

Multiple Examples and Memory Allocation: In the previous sections, it was assumed that a single example image, $\mathbf{x}^e$, was used to learn the MFs. In practice, however, a single image may not deliver the correct properties of the underlying statistics. Hence, it is preferable to learn the MFs from several images. Adding more images into the system can be implemented easily by concatenating all image equations together into a single equation and proceeding as above. If there are R example images denoted $\mathbf{x}^e_1 \cdots \mathbf{x}^e_R$, (24) is extended to

$$L_i\mathbf{p} + \tilde{\mathbf{h}}^e_i = \hat{\mathbf{x}}^e_i \quad \text{for} \quad i = 1 \cdots R$$

where L_i and $\tilde{\mathbf{h}}^e_i$ are calculated as defined above for a single image. The solution $\hat{\mathbf{p}}$ minimizing the LS cost function

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} \left\{ \sum_i^R \|\hat{\mathbf{x}}\left(\mathbf{y}^e_i\right) - \mathbf{x}^e_i\|^2 + \sqrt{\lambda}\|\mathbf{p} - \mathbf{q}\|^2 \right\}$$

gives rise to the following solution which replaces (25):

$$\hat{\mathbf{p}} = \left(\sum_i^R \left({L_i}^T L_i\right) + \lambda I \right)^{-1} \left(\sum_i^R {L_i}^T(\mathbf{x}^e_i - \tilde{\mathbf{h}}^e_i) + \lambda \mathbf{q} \right).$$

Note that the dimensions of ${L_i}^T L_i$ is $K(M+1) \times K(M+1)$ where commonly $K(M+1) \ll N$. Therefore, implementing the above solution needs only a memory capacity on the order of (K^2M^2) which is independent of the number of images. This scheme can also be implemented with a single image if its size is too large. In such a case, the image is split into several subimages, and each subimage is treated as a separate image.

Exploiting the Symmetry of Mapping Functions: The marginal distributions of the wavelet coefficients are known to be symmetric, i.e., $P_i(x^i) = P_i(-x^i)$. This induces symmetric MFs as well. Exploiting this fact, it is possible to limit the SLT interval to include only the positive part of the mapping domain. In this case, the $\mathbf{q}$ values are all positive: $\mathbf{q} = [q_0 \cdots q_M]$ where $q_0 = 0$. The SLT equation $\mathbf{x} = S_{\mathbf{q}}(\mathbf{x})\mathbf{q} + \mathbf{h}$ is still correct if the definition of $S_{\mathbf{q}}(\mathbf{x})$ is modified as follows:

$$[S_{\mathbf{q}}(\mathbf{x})](i,j) = \begin{cases} \operatorname{sign}\{x^i\} r(x^i), & \text{if } \pi(|x^i|) = j \\ \operatorname{sign}\{x^i\}(1 - r(x^i)), & \text{if } \pi(|x^i|) = j+1 \\ 0, & \text{otherwise.} \end{cases}$$

There are two advantages using the new definitions. First, the size of the linear system to be solved is half the size of the original system, enabling more efficiency in memory allocation. Second, pulling more pixel values to the available bins stabilizes the solution and reduces the chance of over-fitting.

Quantization Bins: An important parameter in the proposed scheme is the number of quantization bins used in the SLT. The greater the number of bins used, the more flexibility we gain for the generated MFs (although at the expense of computational burden). It was experimentally demonstrated that relatively few quantization bins (15 bins in our experiment settings) produce

Fig. 3. Images on which the denoising schemes were tested. From left to right top down: Barbara, Boat, Fingerprint, House, Lena, Peppers.

Fig. 4. Images on which the MFs were trained.

superior results that are very close to the asymptotic quality (see results in Section VII). Additionally, since small wavelet values are much more probable that higher values, it is preferable to implement a nonuniform quantization where quantization boundaries are populated more densely in the lower part of the mapping domain. In our experiments, we implemented a polynomial scaling to a set of quantization values $\{t_j\}$ which was produced by uniformly dividing a unit interval, such that each $t_j \in [0,1]$ is mapped to $q_j \in [a,b]$ by applying $q_j = t_j^{\beta} \cdot (b-a) + a$, for $\beta > 1$. It is also possible to develop an iterative scheme for optimal sampling similar to the Lloyd–Max scheme [40]; however, this extension is outside the scope of this paper.

VII. Results

In order to demonstrate the advantages of the proposed approach and to indicate the source of improvements, we compare the denoising results using three different schemes.

- **Method 1** (transform domain—independent bands): A set of MFs is optimized in the transform domain. The optimization is applied to each band independently, minimizing the objective function

$$\varepsilon_k = \left\| \mathbf{x}^e_{\mathrm{B}_k} - \mathcal{M}^k_{\mathrm{B}} \left\{ \mathbf{y}^e_{\mathrm{B}_k} \right\} \right\|^2$$

  where $\mathbf{y}^e_{\mathrm{B}_k} = B_k \mathbf{y}^e$ and similarly $\mathbf{x}^e_{\mathrm{B}_k} = B_k \mathbf{x}^e$. In the case of piecewise linear mapping functions, the above minimization gives

$$\mathbf{p}_k = \left(S_k^T S_k \right)^{-1} S_k^T \mathbf{x}^e_{\mathrm{B}_k} \quad \forall k$$

  where we define $S_k = S_{\mathbf{q}_k}(\mathbf{y}^e_{\mathrm{B}_k})$. Using this method, the solution ignores the statistical dependencies that exist between wavelet coefficients. Note that this minimization criterion is in accord with the traditional shrinkage approaches [5], [8], [9] with the exception that the MFs are optimized here in a LS sense.
- **Method 2** (spatial domain—independent bands): A set of MFs is optimized in the spatial domain. The objective term for this method reads

$$\varepsilon_k = \left\| B_k^T \mathbf{x}^e_{\mathrm{B}_k} - B_k^T \mathcal{M}^k_{\mathrm{B}} \left\{ \mathbf{y}^e_{\mathrm{B}_k} \right\} \right\|^2.$$

  This minimization gives rise to the following solution:

$$\mathbf{p}_k = \left(S_k^T B_k B_k^T S_k \right)^{-1} S_k^T B_k B_k^T \mathbf{x}^e_{\mathrm{B}_k} \quad \forall k.$$

  Note that the objective criterion is expressed in the spatial domain, yet, the MFs are evaluated for each band independently. Thus, while within-band dependencies are considered through the backward projections, interband dependencies are ignored.
- **Method 3** (spatial domain - joint bands): The scheme suggested in this paper (Section V) where the objective goal is expressed in the spatial domain

$$\varepsilon = \left\| \mathbf{x}^e - (B^T B)^{-1} \sum_k B_k^T \mathcal{M}^k_{\mathrm{B}} \left\{ \mathbf{y}^e_{\mathrm{B}_k} \right\} \right\|^2$$

  and the solution is given in (22). In this scheme, the MFs are evaluated simultaneously while taking into account interband as well as intraband dependencies.

It is easy to verify that, for unitary transforms, the three methods listed above eventually coincide and express identical objective functions. This is not the case, however, in the over-complete scheme.

The above methods were tested and compared using a set of experiments. In all the experiments described below, we used the undecimated windowed discrete cosine transform (DCT) as the image transform. Since the undecimated DCT is a tight frame, the term $(B^T B)^{-1}$ in (20) can be ignored, enabling efficient implementation. Note that, due to the undecimated form, each wavelet band can be calculated using a single 2-D separable convolution (with the corresponding DCT basis as the convolution kernel). Additionally, the inverse transform can be applied by convolving the rectified coefficients with the kernels forming B_k^T which are the reflected (180-degree rotation) DCT kernels.

In the following experiments, unless mentioned otherwise, the setting parameters were defined as follows. 1) Test images were taken from Fig. 3. 2) Training was performed on the top-right image of Fig. 4. 3) Transform basis was the undecimated 8×8 DCT. 4) The noise consists of additive Gaussian noise with a s.t.d. of 20 gray levels.

Fig. 5 displays some of the MFs obtained for an 8×8 DCT basis, using the three methods described above. MFs on each row correspond to band indices (i,i) of the 8×8 DCT basis, where $i = 2 \cdot \cdot 6$ (left to right). Note that a DCT band with index

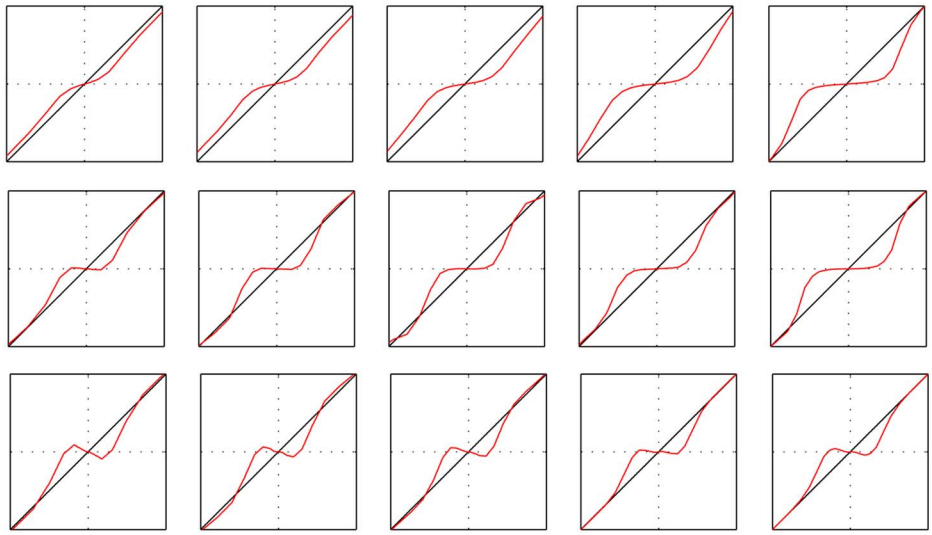

Fig. 5. Comparison of the produced MFs using Method 1 (top row), Method 2 (middle row), and Method 3 (bottom row). MFs on each row correspond to band (i,i) of the 8×8 DCT basis, where $i = 2 \cdot \cdot 6$ (left to right). Graph axes are shown in the range $[-120, 120]$.

(i,j) is a result of convolving the image with a DCT basis whose frequency is i along the x-axis and j along the y-axis. The top and the middle rows show the MFs resulting from the first and the second methods, respectively. It can be seen that these MFs generally resemble the shrinkage shapes of the traditional MFs. The MFs shown in bottom row present the results of the third method. In contrast to the previous methods, here, the MFs produced do not necessarily retain monotonicity and have portions in which positive coefficients are mapped to negative values and portions in which negative coefficients are mapped to positive values, producing regions of negative slope.

The obtained MFs were tested on several images shown in Fig. 3. These images are commonly used as test cases for denoising algorithms.[1] Fig. 6 compares the resulting PSNR for each one of the described methods. The figure is composed of six clusters of bars, each of which compares the denoising results of a particular image. Each bar presents the denoising results averaged over ten realizations of noise with a s.t.d. of 20 gray levels. The results demonstrate the improvement of the second method over the first method, and the superiority of the third method over the other two. Note that the traditional approaches which optimize the MFs in the transform domain are analogous to the first method. It can be seen that most of the improvement is achieved due to formulating the objective in the spatial domain (Method 2). Further improvement, although less significant, is achieved when incorporating the band dependencies into the solution (Method 3). Examples of denoised images after applying Method 3 are shown in Fig. 7.

Running time for the training phase depends, of course, on the size of the example image (or images). In a typical setup, the run time for producing 64 MFs (8×8) DCT basis which were trained on a 1 K $\times$ 1.5 K image was 4.2 min. The program was implemented in Matlab and run on a 1.7-GHz Pentium processor. Applying the 64 MFs to denoising a 512×512 image required 18 s.

[1]Taken from http://decsai.ugr.es/javier/denoise/test_images/index.htm

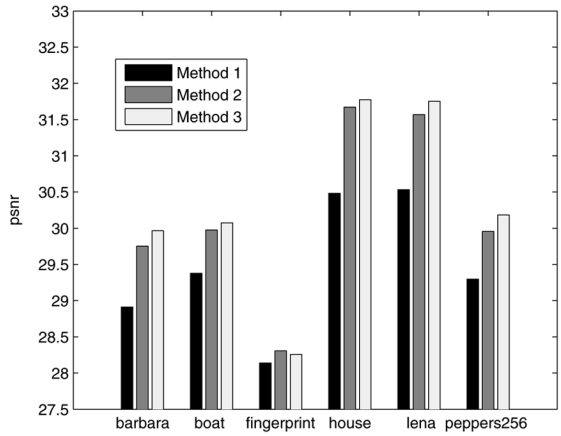

Fig. 6. PSNR after applying the MFs produced by Methods 1–3. Each bar is an average over ten different noise realizations.

A. Nonmonotonic Mapping Functions

The nonmonotonicity and in particular the sign-change in the MFs of Method 3 (Fig. 5 bottom row) are surprising results that were not reported in previous studies. Since this behavior was not observed in Methods 1 and 2, it can be concluded that this phenomena is due to the interband dependencies that are taken into account only in Method 3. An explanation for this behavior is illustrated in Fig. 8. For demonstration purposes, we assume a two-valued signal $\mathbf{x} \in \Re^2$. A signal $\mathbf{x}$ is represented in a unitary transform domain whose bases are the (u,v) axes. Thus, the signal is denoted by a point in the (u,v) plane (see figure). In analogy to the wavelet transform, we assume a signal prior of sparse characteristic in the transform coefficients. Therefore, non-noisy signals are expected to be located within the shaded area extending along the main axes. In this illustration, the true signal $\mathbf{x}$ is located on the u-axis and marked by a white dot. Due to additive noise the acquired signal $\mathbf{y}$ is

Fig. 7. Some examples of denoised images. The images in the top row were contaminated with white noise with a s.t.d. of 20 gray levels. The reconstructed images are shown on the bottom row.

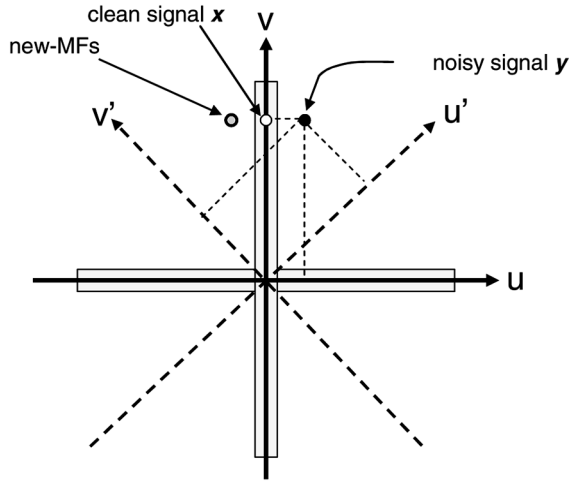

Fig. 8. Two-valued clean signal is represented by the white dot in a unitary transform domain (u, v). The noisy signal is represented by the black dot. An additional unitary transform is represented by the (u', v') axes. Applying the sign-changing MFs to the noisy signal in both transforms (over-complete) results in the clean signal (see text).

measured outside the shaded area denoted by the black dot. Denoising the measured signal using classical shrinkage operations in the (u, v) domain (using, e.g., the hard thresholding MFs) results in a new signal whose u-component is shrunk to zero and the v-component is left untouched (due to its large value). This, indeed, produces the desired solution. Consider now the case where the unitary transform is extended to form a shift-invariant system forming an over-complete transform. An additional unitary transform is appended whose basis vectors are composed of spatially shifted versions of the original transform basis. This additional transform can be viewed in our 2-D example by the (u', v') basis which is obtained by an axes rotation about the origin (due to its unitarity). Applying classical

TABLE I
RESULTING PSNR FOR VARIOUS NOISE LEVELS. THE TRANSFORM USED WAS THE UNDECIMATED 9×9 DCT. THE MFS WERE TRAINED ON THE TOP-RIGHT IMAGE IN FIG. 4

noise s.t.d.	BARB.	BOAT	FGRP.	HOUSE	LENA	PEPP.
1	48.71	48.44	48.41	49.11	48.50	48.46
2	43.69	43.01	42.94	44.40	43.43	43.22
5	38.07	37.00	36.55	39.12	38.48	37.63
10	34.19	33.49	32.27	35.53	35.37	33.84
15	31.95	31.55	29.94	33.52	33.47	31.73
20	30.36	30.19	28.36	32.11	32.10	30.20
25	29.09	29.11	27.15	30.95	31.02	29.04

shrinkage operations to the new transform coefficients provides a signal correction which can handle corrupted signals proximal to the new axes as well (shift invariance). However, this advantage comes with a drawback: Since the two transforms interfere with each other. In some cases, the new estimation is inferior to that obtained with a single unitary transform. In our example, denoising the measured signal in the (u, v) transform produces the white dot, while denoising in the (u', v') transform does not affect the signal (since both coefficients are large enough). Thus, the estimation using the over-complete transform, will result in a signal which is the arithmetic mean of the white and black dots.[2] This estimation is inferior to the previous estimation given by the original unitary transform. The sign change in the MFs as obtained by our numerical optimization solves this problem; The MFs of the (u, v) transform now modify the u-coefficient by *negating* its value (gray dot). The arithmetic mean between the (u, v) transform estimation and the (u', v') estimation (black dot) is now located again within the shaded area. Note that due to symmetry, this behavior is carried out for signals along the (u', v') axes, as well, providing the shift-invariant characteristics. Nevertheless, care must be taken in cases where

[2]It can be easily verified that the pseudo-inverse of a set of unitary transforms is equivalent to the arithmetic mean of the individual transform inverses.

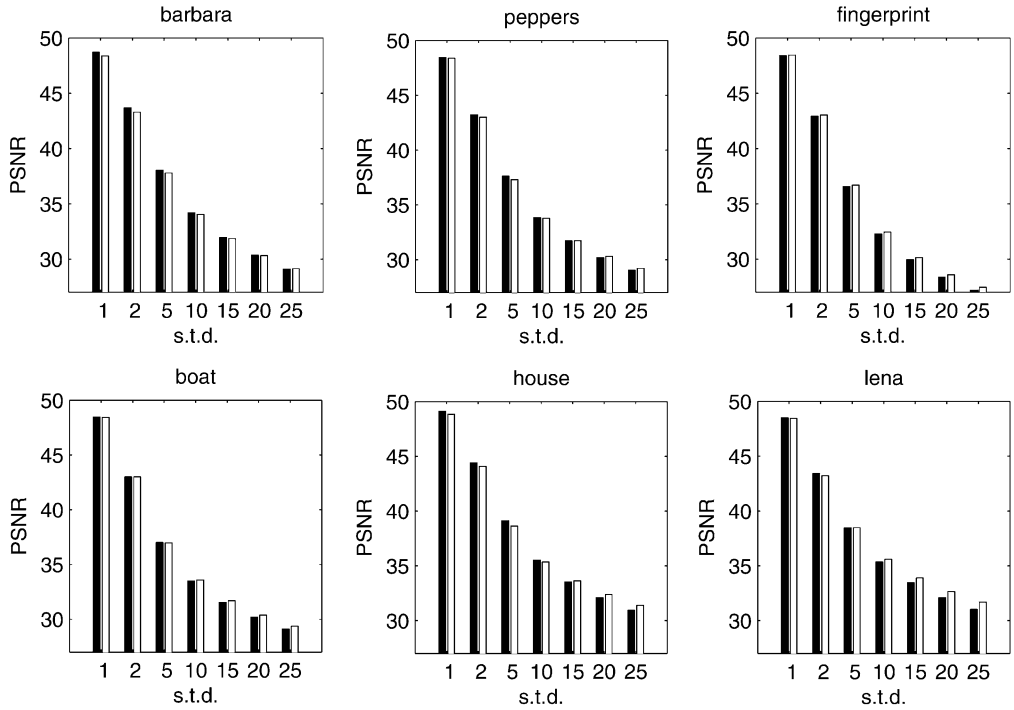

Fig. 9. Comparison between the proposed method and the BLS-GSM method for various noise levels. Dark bars: The porposed method. White bars: The BLS-GSM method.

coefficients include only noise. In such cases, the nonmonotonic adjustment might degenerate the estimation. Since this will be the case mainly for coefficients at finer (high frequency) scales, it appears that the nonmonotonic adjustment is reduced for these bands.

B. Comparison With Other Methods

The proposed approach was tested on the images presented in Fig. 3 which were contaminated with Gaussian white noise under various noise levels. The resulting PSNR are shown in Table I. The transform used in this table was the undecimated 9×9 DCT. Although the transform used is not optimal for natural images and the training image was chosen arbitrarily, the PSNR obtained presents high quality results. These results were compared to the Bayes Least-Squares Gaussian Scale Mixture (BLS-GSM) approach suggested by Portilla *et al.* [25] and considered the state-of-the-art in image denoising. The comparison results are shown in Fig. 9 for each image independently. It is demonstrated that the proposed method presents comparable results with the BLS-GSM method. In low noise variance scenarios the suggested method marginally outperforms BLS-GSM in almost all images, and in more severe noise cases (15 s.t.d. and above) the BLS-GSM demonstrates marginally better performance.

C. Role of Noise Variance

The influence of the noise variance on the obtained MFs can be seen in Fig. 10. Similar to the classical hard/soft thresholding MFs, the profiles of the produced MFs scale down when the noise variance decreases and scale up when the variance increases. The amount of scaling was experimentally shown to follow linearly with the relative increase/decrease in the noise

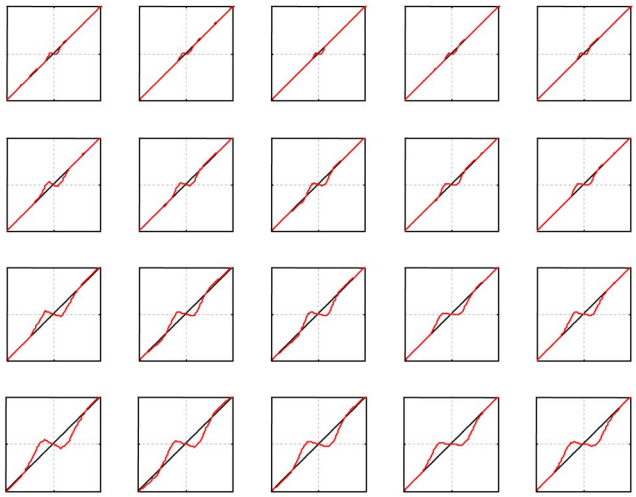

Fig. 10. MFs produced using Method 3 for various noise levels. MFs on each row correspond to band (i,i) of the 8×8 DCT basis, where $i = 2 \cdot \cdot 6$ (left to right). The noise levels were of 5, 10, 15, and 20 s.t.d. from top row to bottom row, respectively. Graph axes are shown in the range $[-120, 120]$.

variance. Thus, if a particular MF $\mathcal{M}_{\sigma_0}\{v\}$ was obtained for noise variance σ_0, the MF for noise variance σ is expected to be

$$\mathcal{M}_{\sigma}\{v\} = s\mathcal{M}_{\sigma_0}\left\{\frac{v}{s}\right\} \quad \text{where} \quad s = \frac{\sigma}{\sigma_0}. \tag{26}$$

This scaling property is very useful as one can estimate the noise variance of a given image using MAD or other available methods (e.g., [37]–[39]) and then apply an appropriately scaled MF set for denoising. An example of two sets of MFs, superimposed on the same plot, one for the $\sigma = 20$ s.t.d. and the second for $\sigma = 10$ s.t.d. scaled by 2, are shown in Fig. 11.

It is demonstrated that the two MFs coincide almost perfectly and are difficult to distinguish. For more extensive experiments demonstrating the scaling relation between the MFs the reader is referred to [31].

D. Training Images

The resemblance between the training images and the target noisy images plays a role in the denoising quality. The influence of this factor is demonstrated in Fig. 6 where the PSNR result of the Fingerprint image is worse for Method 3 than for Method 2. The main reason for this result is that the training image in this experiment (top-right image of Fig. 4) does not seem to be a good representative for the textured Fingerprint image. In order to verify this claim, we tested again the results of Method 3, this time with a training image that is more "similar" to Fingerprint (actually we used the Fingerprintimage rotated by 180°). The results are given in Fig. 12. This plot shows that for all but the Fingerprint image the resulting PSNR are significantly worse, however, for the Fingerprint image, training on a similar textured image exhibits an increase in the resulting PSNR of 0.3 dB.

The left diagram in Fig. 13 presents a set of resulting PSNR using eight different MFs, each of which was trained on a different natural image taken from the set shown in Fig. 4. In this experiment, the choice of the trained natural image influenced the resulting PSNR by up to 0.6 dB. reflecting the role of the training images on the resulting quality. However, this dependency can be significantly reduced by increasing the number of images included in the training set. The diagram in Fig. 13 (right) shows the resulting PSNR using MFs that were trained on several images. The training images were the same images that were used in the left diagram; however, this time the MFs were generated using different numbers of training images ranging from 1 to 8 (left-to-right bars). It is demonstrated that in general the resulting PSNR moderately increases as the number of training images grows. Furthermore, the PSNR fluctuations due to the selection of trained images are drastically reduced.

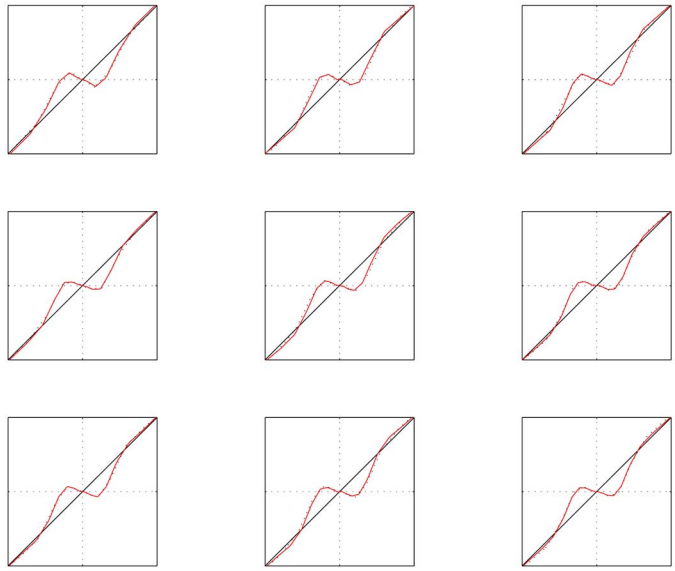

Fig. 11. Comparison between the MFs produced for 20 s.t.d. (red line) and 10 s.t.d. scaled by 2 (black dots). The MFs shown are for DCT bands $[2\cdots4]\times[2\cdots4]$. The graph axes are shown in the range $[-120, 120]$. The two graphs coincide almost perfectly and hard to distinguish.

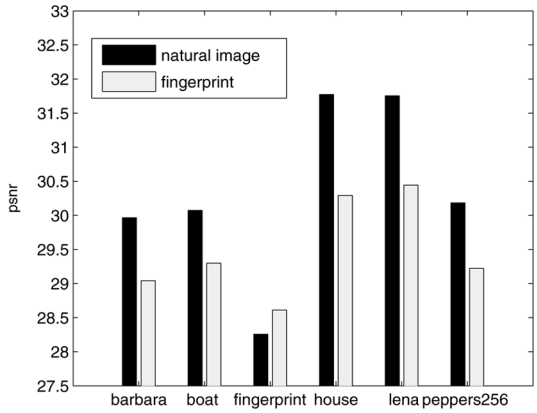

Fig. 12. Comparison results for denoising images using Method 3 where the training images were Fingerprint rotated by 180° (gray bars) and a natural image (black bars).

E. Transform Used

Previous studies demonstrated the benefit of using particular transforms, such as steerable pyramids, curvelets, and contourlet [9], [13], [15] as being more appropriate for modeling natural images. Recent approaches customize the transform used to the noisy image and adaptively learn the transform basis [16]. The scheme presented in this paper is general, and can work with any given transform or any set of filters. In all our experiments, we used the undecimated DCT transform with various window sizes. As it was shown above, the results obtained demonstrate quality comparable with the state-of-the-art methods. It is expected that further improvement can be achieved if other, more appropriate, transforms are used. Fig. 14 presents the denoising results using the undecimated DCT transform with various window sizes. It is shown that the optimal size of the DCT window may vary from image to image. The choice of the most appropriate transform for a given image is still an open problem.

F. Number of Quantization Bins

Fig. 15 shows the resulting PSNR versus the number of quantization bins used. It is shown that about 15 quantization bins are sufficient for high-quality results and that finer quantization does not significantly improve the results. This behavior is a direct outcome of the smooth manner of the optimal MFs. It also strengthens the rationale behind modeling the MFs as piece-wise linear functions. In all other experiments reported in this paper we used 15 quantization bins to define the piece-wise mapping functions. Additionally, since small wavelet values are much more probable than higher values, we implemented a nonuniform quantization as described in Section VI.

G. Other Reconstruction Problems

The approach described in this paper is presented in the context of image denoising, where the contaminated noise is assumed to be Gaussian. However, since the approach does not require any modeling of the image statistics nor of the noise, it

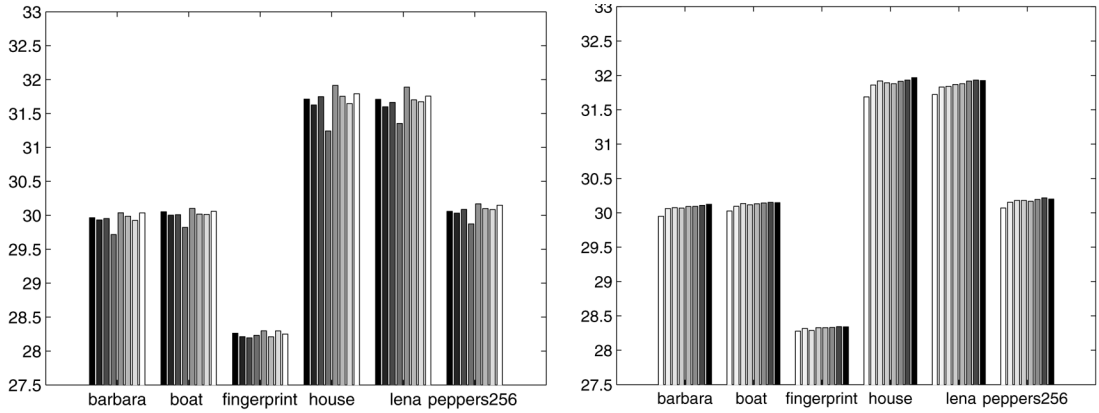

Fig. 13. Left: PSNR of denoised images using MFs that were trained on various natural images shown in Fig. 4. The transform used was DCT 8 × 8. The contaminated noise was Gaussian noise with 20 s.t.d. gray levels. Right: PSNR of denoised images versus the number of training images on which the MFs were trained. Each group of bars shows the PSNR arising from different sized training sets, ranging from 1 ot 8 (left to right).

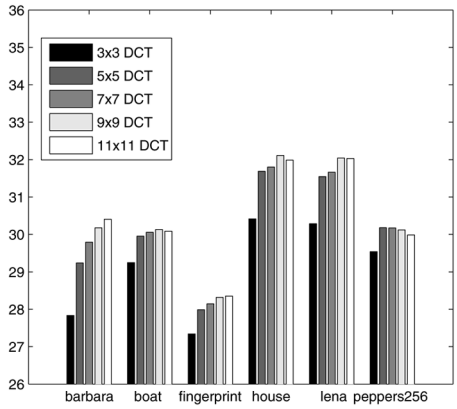

Fig. 14. Resulting PSNR versus DCT window size

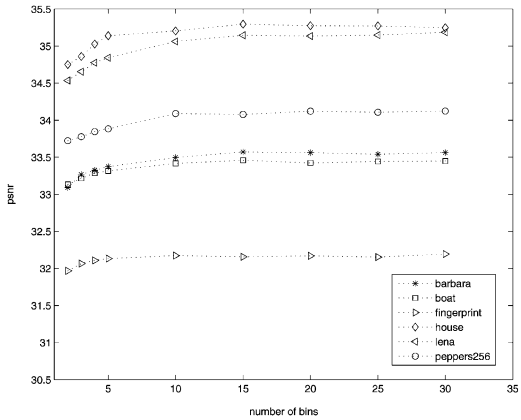

Fig. 15. Resulting PSNR versus the number of quantization bins used for the MFs. The transform used was the undecimated 5 × 5 DCT. The results are shown for various images with noise s.t.d. = 10.

can be seamlessly applied in other reconstruction problems and with different types of noise characteristics. As long as the reconstruction process involves applying scalar look-up-tables in the transform domain, optimal MFs can be obtained. One only needs to provide predegradation and postdegradation images. This section presents some examples of applying the suggested approach to other reconstruction problems, namely: removing JPEG artifacts, and image deblurring. These examples are given in order to demonstrate the concept with no comparative study.

In the first experiment we attempted to deblur images using a set of look-up tables (LUTs) applied to undecimated DCT transform coefficients. The LUTs were trained on the image Lena after it was blurred with a 5-tap Gaussian. A partial set of the produced LUTs are given in Fig. 16 (left). The full set of LUTs were applied to a blurred version of Barbara (same blurring parameters). A close-up view of the deblurred Barbara is given in Fig. 17. The resulting image demonstrates promising sharpening performance with relatively low Gibbs artifacts.

In the second example, the LUTs were trained to reduce severe JPEG artifacts. In this experiment the image Barbara served as the training image and LUTs were applied to Lena. The "noise" was generated by JPEG-compression with quality parameter set to 30%. A partial set of the LUTs are shown in Fig. 16 (right). The JPEG artifacts of the compressed image are presented in Fig. 18 (left) and in a close-up view in Fig. 19 (left). The artifact removal after applying the learned LUTs are shown in Fig. 18 (right) and Fig. 19 (right). The quality of the reconstruction is self-evident. It is interesting to mention the resemblance of the proposed approach to that of Nosratinia [41]. Nosratinia suggested a useful technique for denoising JPEG images by re-applying the JPEG Q-table to shifted versions of the un-compressed image. This technique can be described identically by applying marginal LUTs to the 8 × 8 undecimated-DCT coefficients. In contrast to Nosratinia's approach, the suggested scheme enables the design of a new set of LUTs that are optimized to produce the best results.

VIII. Discussion and Conclusion

This paper suggests a new and simple scheme for wavelet denoising relying on a discriminative framework. One main advantage of the proposed technique is that the shrinkage functions are optimized directly with respect to a set of example images, eliminating the need for modeling complex statistical priors in high-dimensional space. The existence of a statistical prior of natural images is a standard assumption in image processing, and there are several competing models for that prior (e.g., [21],

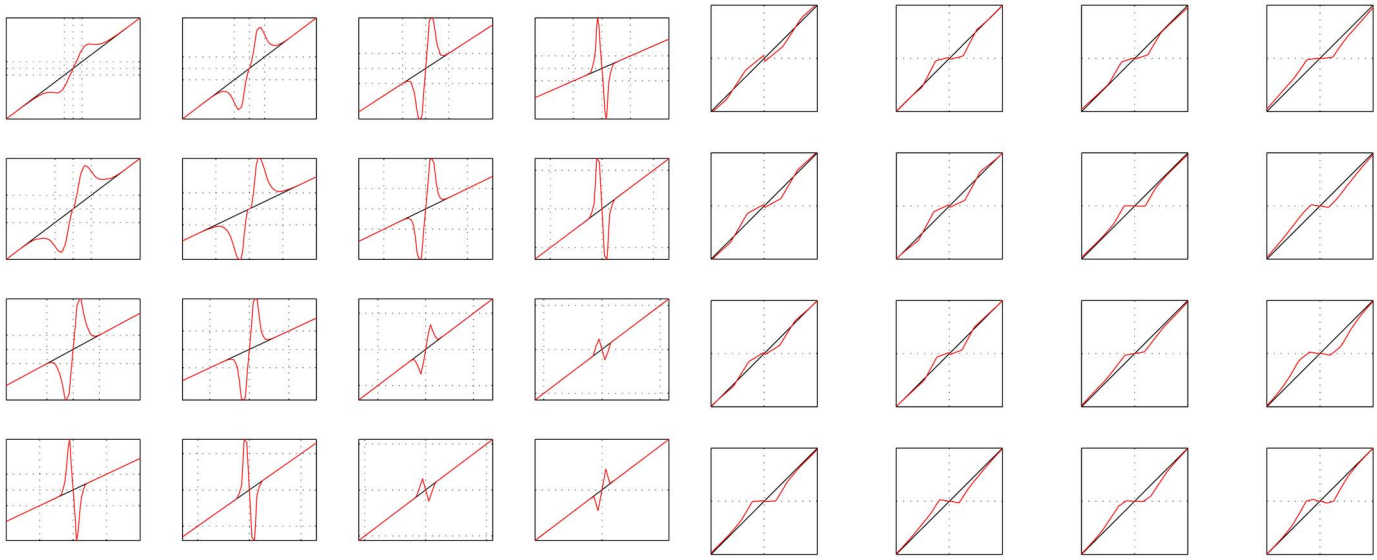

Fig. 16. Left: LUTs that were learned to sharpen images blurred with a 5-tap Gaussain kernel. The filters used were 8×8 DCT. The MFs shown are for the DCT bands whose indices are: $[3 \cdot \cdot 6] \times [3 \cdot \cdot 6]$ (left to right $\times$ top to bottom). The scaling factor of each graph is indicated by the dotted lines, plotted at values $\{-20\ 0\ 20\}$ for each axis. Right: The LUTs that were learned from JPEG-compressed Barbara with $\mathrm{quality} = 30$. The filters used were DCT 8×8. The LUTs shown are for the DCT bands whose indices are $[3 \cdot \cdot 6] \times [3 \cdot \cdot 6]$ (left to right $\times$ top to bottom). Graph axes are shown in the range $[-60, 60]$.

Fig. 17. Left: Blurred Barbara after applying a 5-tap Gaussian blur. Right: Sharpened Barbara using LUTs that were trained on blurred Lena using 8×8 over-complete DCT.

[34], and [42]). Using the suggested scheme, however, we do not need to select between alternative priors, but merely assume a prior exists. Another important generalization in the proposed approach is that there is no need to model the statistical characteristics of the noise, as opposed to most alternatives that typically resort to the easily modeled white Gaussian noise. In contrast, we only assume the existence of a noise model and the technique is applied similarly whether the true noise is simple white Gaussian or more complex (e.g., JPEG noise). Thus, our approach can be applied seamlessly to other degradation processes, as long as the restoration process relies on marginal rectification of transform coefficients. The suggested scheme produces optimal solution with respect to the following aspects.

- An optimal set of scalar MFs (in LS sense) is generated for over-complete transforms taking into account intraband and interband dependencies. As far as we know, previous scalar MF-based techniques ignore these dependencies, as they complicate the statistical models.
- The optimality is expressed in the spatial domain, which is the domain in which the image is perceived. Whereas working in the spatial domain might pose a significant hurdle in the descriptive approach, in the proposed model the restriction to the spatial domain posed only a computational burden.

As emphasized above, the suggested scheme is based on marginal rectification of transform coefficients, namely the MFs are scalar look-up-tables. This restriction is the main limitation of the proposed scheme as possible dependencies on other coefficients cannot be considered adaptively (online). This restriction can be relaxed by applying multivariate MFs, possibly approx-

Fig. 18. Left: JPEG artifacts after compressing Lena with JPEG $\mathrm{quality} = 30$. Right: Artifact removal using LUTs that were trained from JPEG-compressed Barbara using the 8×8 DCT.

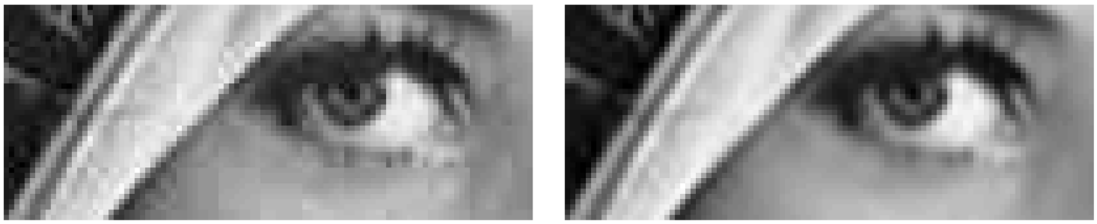

Fig. 19. Left: Zoom-in of Lena JPEG artifacts. Right: Zoom-in of the artifact removal using the proposed method.

imated by embedding quantization bins in higher dimensional spaces. However, since the number of boundary variables increases exponentially with dimensionality, a naive extension is impractical and some sort of dimensionality reduction must be applied. We leave this extension for future work.

Another limitation of the developed scheme is that it relies on the assumption that the noise characteristics are homogeneous. This noise model, although standard in many applications, is imprecise in some real-world scenarios where the noise variance is spatially dependent. An extension of the proposed technique would be to apply an adaptive set of MFs that are scaled adaptively according to the estimated local noise variance.

There are two important issues that were not dealt with in this paper and should be further investigated. The first issue concerns the relation between the transform used and the quality of the denoising results. Clearly, the applied transform plays an important role in the resulting quality (Section VII-E). The transform used should be influenced by the image characteristics as well as the type of contaminating noise. The choice of transform (or set of filters in the case of undecimated transforms) that produces the best results is still an open question.

The second open issue concerns the selection of training images. For simplicity, in this paper, we have arbitrarily chosen natural images for training. This option is reasonable when knowledge about the target images is unknown *a priori*. However, for better results, an attempt should be made to match the test and the training images. Thus, MFs for cartoon type images, for example, should be trained on cartoon examples and MFs trained on a particular texture should be applied to similarly textured images.

ACKNOWLEDGMENT

The authors would like to thank Prof. M. Elad for his helpful comments and stimulating discussions and Prof. J.-L. Starck for providing us benchmark references for denoising algorithms.

REFERENCES

[1] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[2] B. A. Olshausen and D. J. Field, "Natural image statistics and efficient coding," *Network: Comput. Neural Syst.*, vol. 7, pp. 333–339, 1996.

[3] J. Hurri, A. Hyv, R. Karhunen, and E. Oja, "Wavelets and natural image statistics," presented at the Scand. Conf. Image Analysis, Lappenranta, Finland, 1998.

[4] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[5] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, Mar. 1995.

[6] D. L. Donoho and I. M. Johnston, "Ideal denoising in an orthonormal basis chosen from a library of bases," *C. Roy. Acad. Sci.*, vol. 319, pp. 1317–1322, 1994.

[7] D. L. Donoho and I. M. Johnston, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[8] R. R. Coifman and D. L. Donoho, "Translation invariant de-noising," in *Wavelets and Statistics*, A. Antoniadis and G. Oppenheim, Eds. New York: Springer-Verlag, 1995, pp. 125–150.

[9] E. P. Simoncelli and E. H. Adelson, "Noise removal via Bayesian wavelet coring," in *Proc. 3rd Int. Conf. Image Processing*, Lausanne, Switzerland, 1996, vol. I, pp. 379–382.
[10] E. J. Candes, "Harmonic analysis of neural networks," *Appl. Comput. Harmon. Anal.*, vol. 6, pp. 197–218, 1999.
[11] P. Carré and D. Helbert, "Ridgelet decomposition: Discrete implementation and color denoising," presented at the SPIE Wavelet Applications in Industrial Processing III, Boston, MA, Oct. 2005.
[12] N. Nezamoddini-Kachouie, P. Fieguth, and E. Jernigan, "Bayesshrink ridgelets for image denoising," presented at the ICIAR, Porto, Portugal, Sep. 2004.
[13] M. N. Do and M. Vetterli, "The contourlet transform: An efficient directional multiresolution image representation," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2091–2105, Dec. 2005.
[14] B. Matalon, M. Elad, and M. Zibulevsky, "Image denoising with the contourlet transform," presented at the SPARSE, Rennes, France, Nov. 2005.
[15] J. L. Starck, E. J. Candes, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Process.*, vol. 11, no. 6, pp. 670–684, Jun. 2002.
[16] M. Elad and M. Ahron, "Image denoising via sparse and redundant representation over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.
[17] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomoc decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1999.
[18] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
[19] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using hidden markov models," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 886–902, Apr. 1998.
[20] E. P. Simoncelli, "Bayesian denoising of visual images in the wavelet domain," in *Bayesian Inference in Wavelet Based Models*, P. Müller and B. Vidakovic, Eds. New York: Springer-Verlag, 1999, Lecture Notes in Statistics.
[21] E. P. Simoncelli, "Modeling the joint statistics of images in the wavelet domain," in *Proc. SPIE 44th Annu. Meeting*, Denver, CO, 1999, pp. 188–195.
[22] G. Fan and X. Xia, "Image denoising using local contextual hidden markov model in the wavelet domain," *IEEE Signal Process. Lett.*, vol. 8, no. 5, pp. 125–128, May 2001.
[23] A. Pizurica, W. Philips, I. Lemahieu, and M. Acheroy, "A joint inter- and intrascale statistical model for Bayesian wavelet based image denoising," *IEEE Trans. Image Process.*, vol. 11, no. 5, pp. 545–557, May 2002.
[24] S. Chang, B. Yu, and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1522–1531, Sep. 2000.
[25] J. Portilla, V. Strela, M. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.
[26] L. Sendur and I. W. Selesnick, "Bivariate shrinkage with local variance estimation," *IEEE Signal Process. Lett.*, vol. 9, no. 12, pp. 438–441, Dec. 2002.
[27] Z. Shan and S. Aviyente, "Image denoising based on the wavelet co-occurrence matrix," in *Proc. IEEE ICASSP*, Philadelphia, PA, Mar. 2005, pp. 645–648.
[28] X. Li and M. T. Orchard, "Spatially adaptive image denoising under overcomplete expansion," in *Proc. IEEE Int. Conf. Image Processing*, Vancouver, BC, Canada, Mar. 2000, pp. 300–303.
[29] A. Pizurica and W. Philips, "Estimating the probability of the presence of a signal of interest in multiresolution single- and multiband image denoising," *IEEE Trans. Image Process.*, vol. 15, no. 3, pp. 645–665, Mar. 2006.
[30] M. Jansen and A. Bultheel, "Empirical bayes approach to improve wavelet thresholding for image noise reduction," *J. Amer. Statist. Assoc.*, vol. 96, pp. 629–639, 2001.
[31] Y. Hel-Or and D. Shaked, "Slicing the transform—A Discriminative approach for wavelet denoising," Hewlett-Packard Labs Tech. Rep. HPL-2006-103R1, 2006.
[32] D. J. Field, "Wavelets, vision and the statistics of natural scenes," *Phil. Trans. Roy. Soc. London*, vol. 357, no. 1760, 1999.
[33] P. Moulin and J. Liu, "Analysis multiresolution image denoising schemes using generalized-Gaussian priors," in *Proc. IEEE TFTS Symp.*, Pittsburgh, PA, Oct. 6–9, 1998, pp. 633–636.
[34] D. J. Field, "What is the goal of sensory coding," *Neural Comput.*, vol. 6, pp. 559–601, 1994.
[35] A. Antoniadis and J. Fan, "Regularization of wavelet approximation," *J. Amer. Statist. Assoc.*, vol. 96, pp. 939–955, 2001.
[36] M. Elad, "Why simple shrinkage is still relevant for redundant representations?," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5559–5569, Dec. 2006.
[37] P. Kisilev, D. Shaked, and S. H. Lim, "Noise and signal activity maps for better imaging algorithms," presented at the IEEE Int. Conf. Image Processing, 2007.
[38] B. R. Corner, R. M. Narayanan, and S. E. Reichenbach, "Noise estimation in remote sensing imagery using data masking," *Int. J. Remote Sens.*, vol. 24, no. 4, pp. 689–702, Feb. 2003.
[39] K. Rank, M. Lendl, and R. Unbehauen, "Estimation of image noise variance," in *Proc. Vision, Image, and Signal Processing*, 1993, pp. 80–84.
[40] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer.
[41] A. Nosratinia, "Denoising of jpeg images by re-application of jpeg," *J. VLSI Signal Process.*, vol. 27, no. 1, pp. 69–79, 2001.
[42] D. L. Ruderman and W. Bialek, "Statistics of natural images: Scaling in the woods," *Phys. Rev. Lett.*, vol. 73, no. 6, pp. 814–817, 1994.

Yacov Hel-Or received the B.Sc. degree in physics and computer science from Bar-Ilan University, Israel, in 1985, and the Ph.D. degree in computer science from the Hebrew University, Jerusalem, Israel, in 1993.

He is a faculty member at the Interdisciplinary Center, Efi Arazi School of Computer Science, Herzliya, Israel. During 1993 and 1994, he was a Postdoctorate Fellow in the Department of Applied Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot, Israel. From 1994–1996, he was with the NASA Ames Research Center, Moffet Field, CA, as a National Research Council Associate. During 1996–1998, he was a Researcher at the Hewlett Packard Labs, Haifa, Israel. His recent interests include computer vision, image processing, robotics, and computer graphics.

Doron Shaked graduated from the Electrical and Computer Engineering Department, The Ben Gurion University, Beer Sheva, Israel, in 1988, and received the M.Sc. and D.Sc. degrees from the Electrical Engineering Department, The Technion—Israel Institute of Technology, Haifa, Israel, in 1991 and 1995, respectively.

He is a principal Researcher at Hewlett Packard Labs, Haifa, where he has been since 1995, working on color printing and image processing. His other research interests include computer vision, pattern recognition, and shape analysis.

Dr. Shaked was an Ollendorf student fellow (1991) and the recipient of the Wolf prize for excellent students (1994).