Slicing the Transform - A Discriminative Approach for Wavelet Denoising

Yacov Hel-Or, Doron Shaked
HP Laboratories Israel
HPL-2006-103(R.1)
November 8, 2006*

denoising, wavelet, shrinkage, deblurring

This paper suggests a discriminative approach for wavelet denoising where a set of shrinkage functions (SF) are designed to perform optimally (in a MSE sense) with respect to a given set of images. Using the suggested scheme a new set of SFs are generated which are different from the traditional soft/hard thresholding in the over-complete case. These SFs are demonstrated to obtain the state-of-the-art denoising performance. As opposed to the descriptive approaches modeling image or noise priors are not required here and the SFs are learned directly from an ensemble of example images.

* Internal Accession Date Only
Approved for External Publication

Slicing the Transform - 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 shrinkage functions (SF) are designed to perform optimally (in a MSE sense) with respect to a given set of images. Using the suggested scheme a new set of SFs are generated which are different from the traditional soft/hard thresholding in the over-complete case. These SFs are demonstrated to obtain the state-of-the-art denoising performance. As opposed to the descriptive approaches modeling image or noise priors are not required here and the SFs are learned directly from an ensemble of example images.

1 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 post-processing operations. One important field 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 white 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 on 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 usually implemented using some type of wavelet transform. The main motivation for this approach stems from the fact that the wavelet transform of natural images tends to de-correlate pixel values [22, 23, 14, 16]. Hence it is possible to make a reasonable inference on 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 [8, 9, 10]. These techniques amount to modifying the coefficients in the transform domain using a set of scalar mapping functions, $\{\mathcal{M}^i\}$, called *shrinkage functions* (SF). The shrinkage approach comprises of a wavelet transform:

$$\mathbf{y}_{\mathrm{w}} = W\mathbf{y}$$

followed by a correction step in which the wavelet coefficients are rectified according to a set of shrinkage functions (SFs):

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

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

$$\hat{\mathbf{x}} = W^{-1}\hat{\mathbf{x}}_{\mathrm{w}}$$

This process is outlined in Figure 1.

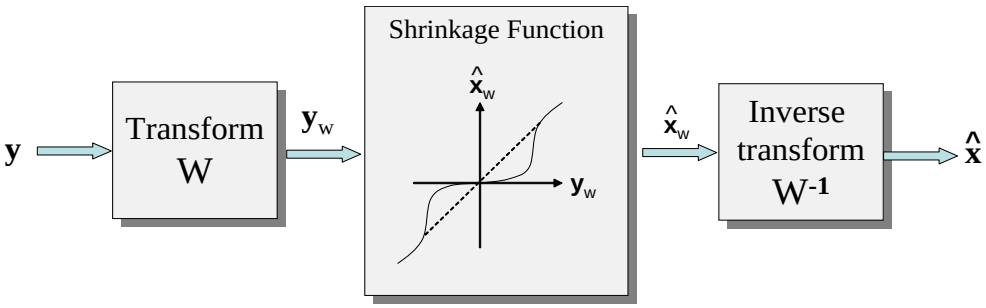

Figure 1: The pipe-line of Wavelet shrinkage.

Due to their simplicity and good results, shrinkage approaches 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 SFs are applied. The justification of applying a marginal (scalar) SF to each coefficient independently can be shown to emerge from the independence assumption of the wavelet coefficients. This assumption was postulated in the early studies in which SFs were applied to unitary transforms.

Since the pioneering work of Donoho and Johnston various efforts have been made to improve the denoising results of shrinkage methods. Such efforts generally concentrated on two main approaches. The first approach attempts to improves the results by abandoning the

unitary representation and working in over-complete transform domains. Such transforms include the un-decimated wavelets [5], steerable wavelets [31], and other recently developed transforms such as ridgelets [1, 2, 20], contourlets [7, 18], and curvelets [32]. These transforms were shown to better represent natural images in the sense that their coefficients tend to exhibit better compaction (sparsity). Additionally, the over-completeness was shown to significantly improve denoising performance. Although the independence assumption cannot be justified in the over-complete domain, most of the conventional methods naively borrowed the traditional SFs from the unitary case.

The second approach toward improvement relaxed the independence assumption of the wavelet coefficients and concentrated on modeling the statistical dependencies between neighboring coefficients. This scheme can be seen as diverging from the scalar SFs to multivariate SFs where transform coefficients are rectified according to a group of measured coefficients. Inter-coefficient dependencies are exploited using any of a range of techniques such as: joint sparsity assumption [4, 17], HMM and Bayesian models [6, 29, 30, 12, 24], context modeling [3, 25], tree models representing parent-child dependencies [27], co-occurrence matrix [28], adaptive thresholding [15], and more. These types of techniques sometimes 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 shrinkage functions (SFs), regardless of the approach used, is that the SFs 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 SFs are derived (scalar or multivariate, parametric or non-parametric) designed to rectify the contaminated coefficients. Clearly, imprecise modeling of the statistical prior leads directly to a deterioration in the resulting performance. Because inter-coefficient dependencies are complicated to model, in particular in the over-complete case, it is expected that the statistical models are far from being precise. And indeed due to the high dimensionality of the joint probability, ad-hoc assumptions commonly have been made in order to make the problem tractable. Such assumptions include, e.g., ignoring the inter-coefficient dependencies (e.g. [31, 8]), modeling only bivariate or parent-child dependencies (e.g. [27]), and modeling the joint dependencies of a small group of neighboring coefficients but assuming simplified parametric models (e.g. [25]).

This paper suggests, *inter alia*, a new technique for designing a set of SFs 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 SFs is constructed using an ensemble of example images whose clean and contaminated versions are supplied off-line. The SFs 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 SFs that are applied to each wavelet

coefficient independently. Nevertheless, although the SFs are applied in a marginal manner, their construction is affected by inter-coefficient dependencies. In fact, the obtained SFs differ from the conventional monotonic functions. Moreover, despite the fact that scalar SFs 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 sources:

- First, the SFs are constructed in an optimal manner taking into account inter-coefficient dependencies. Although the SFs are non-linear their construction is performed in a closed form solution using an image representation which is referred to as: *The Slicing Transform* (SLT). The SLT is a spline based image representation in which non-linear mapping operations can be applied linearly. This property permits the optimal set of SFs to be identified as a solution of a least-squares problem.

- The second source of improvement stems from the optimality criteria applied in this method. While most shrinkage approaches optimize the solution with respect to the MAP criterion, the proposed method is optimized with respect to the minimum squared error (MSE). Because the MAP solution considers only the most probable case, it is possible that the method performs poorly with cases that are not the most probable but are likely. This leads to performance attenuation in the average case. The MAP objective is commonly adopted in denoising techniques due to its mathematical simplicity. However, using the discriminative approach, as proposed here, the MSE objective can be applied efficiently.

- 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 specified with respect to the spatial domain which is the domain where 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 is not properly extended to over-complete transforms. Rather, it can be shown that the optimal solution in the over-complete transform domain does not guarantee optimality in the spatial domain.

The approach suggested in this paper is presented in the context of denoising. However, this technique goes beyond the noise reduction problem and can be applied in a similar manner to other reconstruction problems, such as image de-blurring, image up-scaling, etc. as long as the reconstruction process involves scalar look-up-tables applied in the wavelet domain. Some preliminary 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 4, the *slicing transform* is introduced along with its appealing properties. Section 5 presents the proposed method, and Section 6 addresses several computational issues. Simulation results as well as implementations in other restoration problems are presented in Section 7.

2 Image Restoration in Unitary Transform Domains

The justification for using scalar mapping functions as the SFs can be shown to emerge from the MAP estimation and the independence assumption of the wavelet coefficients. Consider a degradation model as described in Eq. 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 the wavelet transforms $\mathbf{x}_{\mathrm{w}} = W\mathbf{x}$ and $\mathbf{y}_{\mathrm{w}} = W\mathbf{y}$, the MAP estimation gives:

$$\hat{\mathbf{x}}_{\mathrm{w}}(\mathbf{y}_{\mathrm{w}}) = \arg\max_{\mathbf{x}_{\mathrm{W}}} P(\mathbf{x}_{\mathrm{w}}|\mathbf{y}_{\mathrm{w}}) \tag{2}$$

Using the Bayes conditional rule and exploiting the monotonicity of the *log* function, the maximization in Eq. 2 is equivalent to:

$$\hat{\mathbf{x}}_{\mathrm{w}} = \arg\max_{\mathbf{x}_{\mathrm{W}}} P(\mathbf{y}_{\mathrm{w}}|\mathbf{x}_{\mathrm{w}})P(\mathbf{x}_{\mathrm{w}}) = \arg\min_{\mathbf{x}_{\mathrm{W}}} \{-\log P(\mathbf{y}_{\mathrm{w}}|\mathbf{x}_{\mathrm{w}}) - \log P(\mathbf{x}_{\mathrm{w}})\} \tag{3}$$

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

$$-\log P(\mathbf{y}_{\mathrm{w}}|\mathbf{x}_{\mathrm{w}}) = \lambda\|\mathbf{x}_{\mathrm{w}} - \mathbf{y}_{\mathrm{w}}\|^2 = \lambda\sum_i \|x^i_{\mathrm{w}} - y^i_{\mathrm{w}}\|^2 \tag{4}$$

where x^i_{w} and y^i_{w} denote the ith elements of the corresponding vectors and λ is a constant depending on the noise variance. The second term in Equation 3, $\log P(\mathbf{x}_{\mathrm{w}})$, 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:

$$-\log P(\mathbf{x}_{\mathrm{w}}) = -\log\prod_i P_i(x^i_{\mathrm{w}}) = -\sum_i \log P_i(x^i_{\mathrm{w}}) \tag{5}$$

Substituting Equations 4 and 5 into Equation 3 the overall minimization amounts to a set of independent scalar minimizations each of which corresponds to a particular coefficient:

$$\hat{x}^i_{\mathrm{w}}(y^i_{\mathrm{w}}) = \arg\min_{x^i_W}\left\{\lambda\|x^i_{\mathrm{w}} - y^i_{\mathrm{w}}\|^2 - \log P_i(x^i_{\mathrm{w}})\right\} \quad \forall\ i \tag{6}$$

The last expression gives the justification for applying a scalar SF to each wavelet coefficient independently: Each value y^i_{W} is mapped to: $\hat{x}^i_{\mathrm{W}} = \mathcal{M}^i_{\mathrm{W}}\{y^i_{\mathrm{W}}\}$ which is given in Eq. 6. Note, that for a particular noise variance, the variations in the SFs $\mathcal{M}^i_{\mathrm{W}}\{\cdot\}$ depend solely on $P_i(x^i_{\mathrm{W}})$. Furthermore, assuming the statistics of natural images are homogeneous it implies that all wavelet coefficients belonging to a particular wavelet band share the same distribution. Namely, w.l.o.g. if a coefficient x^i_{W} belongs to the j^{th} band where $j = band(i)$, we have:

$$P_i(x^i_{\mathrm{W}}) = P_{band(i)}(x^i_{\mathrm{W}})$$

and

$$\hat{x}^i_{\mathrm{W}} = \mathcal{M}^{band(i)}_{\mathrm{W}}\{y^i_{\mathrm{W}}\} \tag{7}$$

As a result, if the wavelet transform is composed of K bands, only K distinct SFs must be evaluated.

There is a wealth of papers dealing with the estimation of the SFs in the context of denoising. The early studies of Donoho and Johnston suggested using *soft thresholding* or *hard thresholding* as shrinkage functions [9, 10]. 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 Laplacian distribution (i.e. $p = 1$) while hard-thresholding assumes a sharper distribution with $p = 0.5$ [29, 19]. Later studies extended the thresholding approach to other values of p, by applying parameter estimation to the measured coefficients [19, 31, 13]. Other studies implement non-parametric representations for the marginal distributions and calculate the SFs using numerical methods [29].

3 Restoration in Over-Complete Domains

Although the shrinkage approach using unitary transforms provides good results, significant improvement is achieved when implementing this technique using over-complete representations. In most cases, this is implemented using the un-decimated wavelet transform or any other sliding windowed transforms (sliding local DCT, sliding local DFT, steerable pyramid, etc.).

The un-decimated wavelet transform can be viewed as applying an orthogonal transform to a set of shifted versions of the image. The shrinkage operation is applied to each transformed image independently. This is followed by transforming each modified transform back to the image space, and averaging all the corrected images after shifting them back to their original positions. This is illustrated in Figure 2. This procedure was first suggested by Coifman and Donoho where they termed it *cycle-spinning* denoising [5].

The cycle-spinning approach can also be viewed as a single shrinkage operation applied to an over-complete transform. Since the transform of a spatially shifted image can be

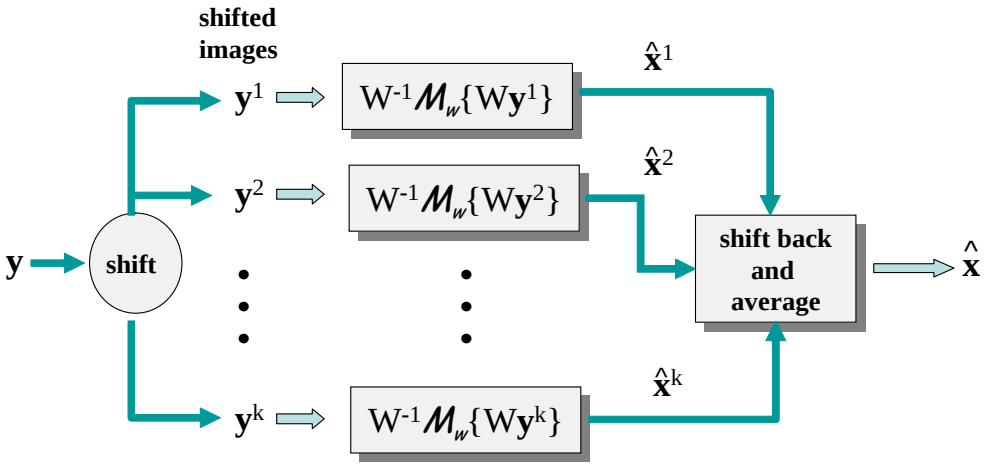

Figure 2: The pipe-line of the cycle-spinning approach.

applied equivalently by shifting the transform basis (by the same amount but in the opposite direction), it is possible to define a set of transforms:

$$\mathbf{y}_{\mathrm{w}_i} = WS_i\mathbf{y} = W_i\mathbf{y} \quad i = 1 \ldots N$$

where $S_i\mathbf{y}$ denotes the i^{th} shift of an image $\mathbf{y}$, and the matrix $W_i = WS_i$ is composed of the wavelet basis after applying the respective shift. Given a set of different shifts, the entire transform gives:

$$\mathbf{y}_{\mathrm{w}} = W\mathbf{y} \tag{8}$$

where now

$$W = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_N \end{bmatrix} \quad \text{and} \quad \mathbf{y}_{\mathrm{w}} = \begin{bmatrix} \mathbf{y}_{\mathrm{w}_1} \\ \mathbf{y}_{\mathrm{w}_2} \\ \vdots \\ \mathbf{y}_{\mathrm{w}_N} \end{bmatrix} \tag{9}$$

Modifying $\mathbf{y}_{\mathrm{w}}$ using a vector of SFs $\vec{\mathcal{M}}_{\mathrm{w}}\{\mathbf{y}_{\mathrm{w}}\}$ aims at removing the noise components. Hence, it is assumed that:

$$W\mathbf{x} = \vec{\mathcal{M}}_{\mathrm{w}}\{W\mathbf{y}\}$$

and the LS estimate for $\mathbf{x}$ gives:

$$\hat{\mathbf{x}} = \left(W^TW\right)^{-1}W^T\vec{\mathcal{M}}_{\mathrm{w}}\{W\mathbf{y}\} \tag{10}$$

In the un-decimated wavelet case, however, Equation 10 can be simplified due to the fact that W is a tight frame, namely

$$\frac{1}{N}W^TW = \frac{1}{N}\sum_i W_i^TW_i = I$$

Additionally, assuming that the SFs are identical for each wavelet transform W_i, i.e: $\vec{\mathcal{M}}_{\mathrm{W}_i} = \vec{\mathcal{M}}_{\mathrm{W}}, \quad \forall\, i = 1..N$, Equation 10 can be rewritten:

$$\hat{\mathbf{x}} = \frac{1}{N} W^T \vec{\mathcal{M}}_{\mathrm{W}} \{W\mathbf{y}\} = \frac{1}{N} \sum_i W_i^T \vec{\mathcal{M}}_{\mathrm{W}} \{W_i \mathbf{y}\} \tag{11}$$

Equation 11 spells out the cycle-spinning procedure as described in Figure 2.

Viewing the cycle-spinning approach as a set of independent shrinkage operations applied to a set of unitary transforms may suggest that the SFs applied are similar to those applied in a single unitary case (Eq. 6). Nevertheless, this is not the case. Considering the entire overcomplete transform (Eq. 8) makes it clear that even if we may assume statistical independence in the coefficients $\mathbf{y}_{\mathrm{W}_i}$ belonging to a single transform W_i, this definitely can not be extended to coefficients belonging to different transforms, e.g. $\mathbf{y}_{\mathrm{W}_i}$ and $\mathbf{y}_{\mathrm{W}_j}$, $i \neq j$. Thus, a new set of SFs should be designed that takes into consideration the inter-transform 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 SFs for the unitary case with respect to the MSE criterion. Namely, finding $\vec{\mathcal{M}}_{\mathrm{W}}$ that minimizes

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

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

$$E_{\mathbf{x}|\mathbf{y}}\{\|W^T \vec{\mathcal{M}}_{\mathrm{W}}\{W\mathbf{y}\} - \mathbf{x}\|^2\} = E_{\mathbf{x}|\mathbf{y}}\{\|\vec{\mathcal{M}}_{\mathrm{W}}\{\mathbf{y}_{\mathrm{W}}\} - \mathbf{x}_{\mathrm{W}}\|^2\} \tag{12}$$

However, for an over-complete transform this equality in not valid anymore (see Appendix A) which implies that the optimization for $\vec{\mathcal{M}}_{\mathrm{W}}$ 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 SFs may not be longer justified when the transform coefficients are mutually dependant, the superior results of applying scalar SFs in the overcomplete case suggest that such a scheme is still very useful in addition to its appealing efficiency. Furthermore, in a recent paper Elad [11] justifies scalar SFs as being the 1^{st} step in an iterative minimization scheme. Justified as the optimal solution or not, there is definitely an interest in finding the best SFs in the over complete case while considering inter-coefficient dependencies. To the best of our knowledge, the optimal design of SFs in the overcomplete domain was not discussed in the literature, and in most cases the applied SFs were naively borrowed from the unitary case.

In this paper, we present a new scheme for image denoising in the over-complete case. The new scheme uses a *discriminative* framework in which the SFs are calculated directly from

a set of example images, overriding the demanding efforts of modeling statistical priors in high-dimensional spaces. As an introduction to the proposed approach we first introduce the *Slicing Transform* (SLT). The SLT will be used in later sections to calculate the optimal SFs.

4 The Slicing Transform and its Properties

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

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

such that

$$q_0 = a < q_1 < q_2 \ldots < \quad q_M = b$$

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

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

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 $q_{\pi(x)-1}$:

$$x = r(x)q_{\pi(x)} + (1 - r(x))q_{\pi(x)-1} \tag{13}$$

Eq. 13 can be rewritten in vectorial form:

$$x = S_{\mathbf{q}}(x)\mathbf{q} \tag{14}$$

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

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

and where the values $1 - r(x)$ and $r(x)$ are located in the $(\pi(x) - 1)^{th}$ and $\pi(x)^{th}$ entries, respectively.

We now define a vectorial extension of Eq. 14. Let $\mathbf{x}$ be an N dimensional vector whose elements satisfy $x^i \in [a, b)$. The *Slicing Transform* (SLT) of $\mathbf{x}$ is defined as follows:

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

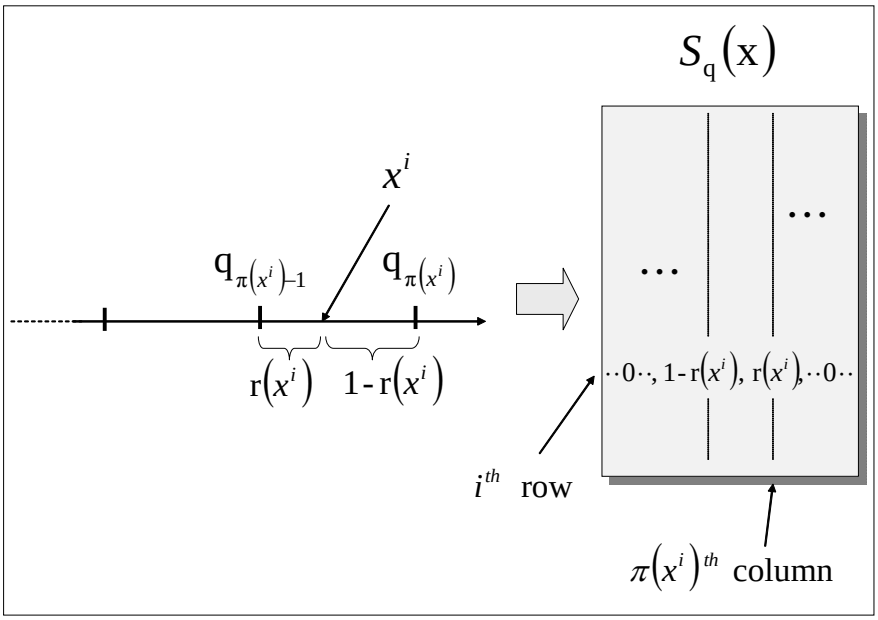

Figure 3: Representing x as a quantized value and residue.

where $S_{\mathbf{q}}(\mathbf{x})$ is an $N \times (M+1)$ matrix defined as follows (see Figure 3):

$$[S_{\mathbf{q}}(\mathbf{x})](i,j) = \begin{cases} r(x^i) & \text{if } \pi(x^i) = j \\ 1 - r(x^i) & \text{if } \pi(x^i) = j+1 \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

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 defined in $\mathbf{q}$. Since every x^i has two adjacent bin boundaries, there are exactly two non-zero values in each row of $S_{\mathbf{q}}(\mathbf{x})$ (with the exception of values x^i equaling a bin boundary value q_j). Correspondingly, in each column j the non-zero entries are only in rows i for which $\pi(x^i) = j$ or $\pi(x^i) = j+1$ (i.e. the values $\{x^i\}$ are adjacent to the bin boundary q_j).

The SLT (Eq. 15) is eventually nothing but a spline representation of a signal $\mathbf{x}$. It can be regarded as a linear transform where the signal $\mathbf{x}$ is represented as a linear combination of column vectors of matrix $S_{\mathbf{q}}(\mathbf{x})$. However, unlike traditional transforms where the basis vectors are fixed and the linear combinations vary according to a given signal $\mathbf{x}$, in the SLT the linear combination weights $\mathbf{q}$ are fixed while the transform matrix varies according to the signal. Further extending the analogy to traditional transforms, some operations that are complicated to apply in the signal domain, may be applied efficiently in the SLT domain. This point will be clear in the paragraphs below.

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 $j \in \{1, 2, \ldots M\}$ 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 Figure 4 for an illustration of such a mapping).

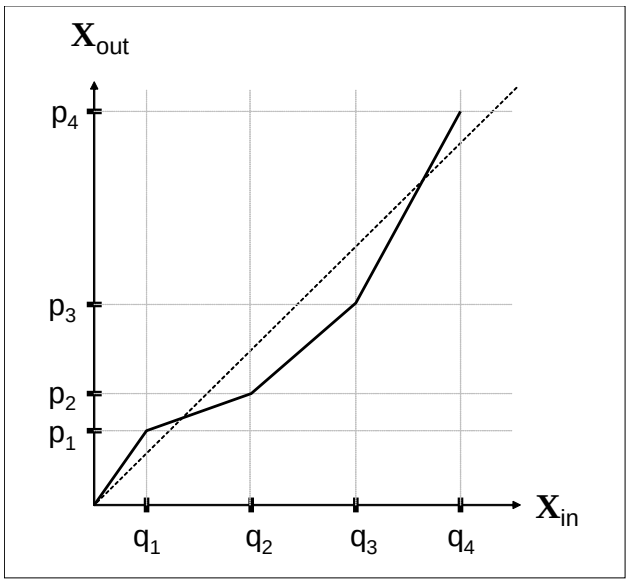

Figure 4: Illustration of a piecewise linear map in accordance with the substitution property of the SLT.

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

In addition to the substitution property, due to the linear structure of the SLT any linear operation $\mathcal{L}$ applied to $\mathbf{x}$ can be applied directly to the SLT representation. This means:

$$\mathcal{L}\{\mathbf{x}\} = \mathcal{L}\{S_{\mathbf{q}}(\mathbf{x})\mathbf{q}\} = \mathcal{L}\{S_{\mathbf{q}}(\mathbf{x})\}\mathbf{q}$$

and consequently:

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

Note however, that once a linear operation has been applied to the SLT, the linear representation loses the SLT form, and thus generally we have that:

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

5 Image Restoration using the SLT

Considering again the restoration scheme in the over-compete domain (Eq. 8) where we have:

$$\mathbf{y}_{\mathrm{w}} = W\mathbf{y}$$

The rows of W are composed of the analysis vectors, the number of which is equal or larger than the dimensionality of the signal $\mathbf{y}$. We recall that our main goal is to find a vector of SFs: $\vec{\mathcal{M}}_{\mathrm{w}} = [\mathcal{M}^1_{\mathrm{w}}, \mathcal{M}^2_{\mathrm{w}}, \cdots]$ that would best restore $\mathbf{x}$ from $\mathbf{y}_{\mathrm{w}}$ (Eq. 10). To simplify the explanations we first present the case where a single SF: $\mathcal{M}_{\mathrm{w}} = \mathcal{M}^i_{\mathrm{w}},\ \forall i$, is used for all coefficients $\mathbf{y}_{\mathrm{w}}$. Later we extend the proposed approach to include a vector of SFs. Recall, that the modified coefficients $\mathcal{M}_{\mathrm{w}}\{\mathbf{y}_{\mathrm{w}}\}$ are designed to restore the original signal $\mathbf{x}$, thus it is assumed that (Eq. 10):

$$\hat{\mathbf{x}}(\mathbf{y}) = (W^TW)^{-1}W^T\mathcal{M}_{\mathrm{w}}\{\mathbf{y}_{\mathrm{w}}\} \tag{17}$$

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

$$\mathbf{y}_{\mathrm{w}} = S_{\mathbf{q}}(\mathbf{y}_{\mathrm{w}})\mathbf{q} \quad \text{and} \quad \mathcal{M}_{\mathbf{q},\mathbf{p}}\{\mathbf{y}_{\mathrm{w}}\} = S_{\mathbf{q}}(\mathbf{y}_{\mathrm{w}})\mathbf{p} \tag{18}$$

where $\mathcal{M}_{\mathbf{q},\mathbf{p}}$ describes the piecewise linear approximation of the mapping $\mathcal{M}_{\mathrm{w}}$. Using Eq. 17, the estimated signal is then given by:

$$\hat{\mathbf{x}}(\mathbf{y}) = (W^TW)^{-1}W^TS_{\mathbf{q}}(\mathbf{y}_{\mathrm{w}})\mathbf{p} \tag{19}$$

The off-line step of the proposed scheme aims at learning the optimal SF to be applied. Namely, the goal is to find the optimal $\mathbf{p}$ vector, that together with the $\mathbf{q}$ vector define the piecewise mapping function $\mathcal{M}_{\mathbf{q},\mathbf{p}}$. The optimal (piecewise) SF with respect to the MSE criterion is obtained by a vector $\hat{\mathbf{p}}$ satisfying:

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} E_{\mathbf{x}|\mathbf{y}}\left\{\|\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}\|^2\right\} \tag{20}$$

where $\hat{\mathbf{x}}(\mathbf{y})$ is as given in Eq. 19. Since the conditional probability $P(\mathbf{x}|\mathbf{y})$ is not available and is complicated to model we approximate the MSE expression using a set of clean signals $\{\mathbf{x}^e\}$ that are given along with their noisy versions $\{\mathbf{y}^e\}$. For simplicity, we first assume that a single signal $\mathbf{x}^e$ is given as an example along with its noisy version $\mathbf{y}^e$. In such a case, the MSE solution is approximated by the following minimization:

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

Claim 1 *The optimal piecewise linear mapping function as defined in Equation 21 is given by:*

$$\hat{\mathbf{p}} = \left(L^T L\right)^{-1} L^T \mathbf{x}^e \tag{22}$$

where

$$L = (W^T W)^{-1} W^T S_{\mathbf{q}}(\mathbf{y}_{\mathrm{W}}) \tag{23}$$

Proof:
According to Equations 19 and 23

$$\hat{\mathbf{x}}(\mathbf{y}^e) = L\mathbf{p}$$

Substituting the above into Eq. 21 gives rise to

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} \|L\mathbf{p} - \mathbf{x}^e\|^2$$

Taking the derivative of the above equation with respect to $\mathbf{p}$ and equating it to zero yields the solution in Eq. 22 $\square$

In fact, the SF in Eq. 22 is designed to optimally reconstruct the clean example from its noisy counterpart. The obtained SF is then used for denoising new signals that are assumed to have similar statistical characteristics as the training example (signal and noise). Clean and noisy examples can be acquired prior and following to the degradation process (e.g. before and after a noisy channel, before and after JPEG compression, etc.). If clean versions of noisy signals are not readily available, then the degradation process must be modeled and noisy signals are simulated from clean examples.

We now extend the solution to the case where a vector of SFs is sought. Considering again the un-decimated wavelet transform (Eq. 8) and the insight (Eq. 7) that mapping functions applied to wavelet coefficients belonging to the same wavelet band are identical. We initially reorder the rows of the over complete transform W in Eq. 9 so that transform rows corresponding to a wavelet band are co-located in a block. Naturally, we extend the same reordering to $\mathbf{y}_{\mathrm{W}}$. Assuming we have K different wavelet bands and a corresponding permutation matrix P:

$$B = PW = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_K \end{bmatrix} \quad \text{and} \quad \mathbf{y}_{\mathrm{B}} = PW\mathbf{y} = P\mathbf{y}_{\mathrm{W}} = \begin{bmatrix} \mathbf{y}_{\mathrm{B}_1} \\ \mathbf{y}_{\mathrm{B}_2} \\ \vdots \\ \mathbf{y}_{\mathrm{B}_K} \end{bmatrix} \tag{24}$$

where $\mathbf{y}_{\mathrm{B}_i}$ represents the coefficients in the i^{th} band. In the new reordering a vector of SFs, $\vec{\mathcal{M}}_{\mathrm{B}} = [\mathcal{M}^1_{\mathrm{B}}, \mathcal{M}^2_{\mathrm{B}}, \cdots, \mathcal{M}^K_{\mathrm{B}}]$, is applied as follows: Since $\mathcal{M}^k_{\mathrm{B}}$ is applied individually to all coefficients in the k^{th} band, we can rewrite Eq. 17 as

$$\hat{\mathbf{x}} = (B^T B)^{-1} B^T \vec{\mathcal{M}}_{\mathrm{B}}\{\mathbf{y}_{\mathrm{B}}\} = (B^T B)^{-1} \sum_{k=1}^{K} B_k^T \mathcal{M}^k_{\mathrm{B}}\{\mathbf{y}_{\mathrm{B}_k}\} \tag{25}$$

where we have:

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

Note, that in contrast to the above case (Eq. 17), different SFs are applied to different $\mathbf{y}_{\mathrm{B}_k}$. Here again we restrict our SFs to be piecewise linear, where:

$$\mathbf{y}_{\mathrm{B}_k} = S_{\mathbf{q}_k}(\mathbf{y}_{\mathrm{B}_k})\mathbf{q}_k \quad \text{and} \quad \mathcal{M}_{\mathbf{q}_k,\mathbf{p}_k}(\mathbf{y}_{\mathrm{B}_k}) = S_{\mathbf{q}_k}(\mathbf{y}_{\mathrm{B}_k})\mathbf{p}_k$$

and $\mathcal{M}_{\mathbf{q}_k,\mathbf{p}_k}$ describes the piecewise linear approximation of the mapping $\mathcal{M}_{\mathrm{B}}^k$. Whereby we can rewrite Eq. 25 as

$$\hat{\mathbf{x}}(\mathbf{y}) = (B^T B)^{-1} \sum_{k=1}^{K} B_k^T S_{\mathbf{q}_k}(\mathbf{y}_{\mathrm{B}_k})\mathbf{p}_k \tag{26}$$

Similar to the above, having $\mathbf{x}^e$ and $\mathbf{y}^e$ as examples, we are looking for the optimal $\mathbf{p}$ that will minimize:

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} \|\hat{\mathbf{x}}(\mathbf{y}^e) - \mathbf{x}^e\|^2 \tag{27}$$

but now $\mathbf{p}$ is a $K(M+1)$ dimensional vector constructed by stacking all $\mathbf{p}_k$:

$$\mathbf{p} = \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_K \end{bmatrix}$$

Claim 2 *The optimal set of piecewise linear mapping functions that minimizes Equation 27 is given by:*

$$\hat{\mathbf{p}} = \left(\tilde{L}^T \tilde{L}\right)^{-1} \tilde{L}^T \mathbf{x}^e \tag{28}$$

where $\tilde{L}$ is composed of all the slicing matrices:

$$\tilde{L} = (B^T B)^{-1} \begin{bmatrix} B_1^T S_{\mathbf{q}_1}(\mathbf{y}^e_{\mathrm{B}_1}) & B_2^T S_{\mathbf{q}_2}(\mathbf{y}^e_{\mathrm{B}_2}) & \cdots & B_K^T S_{\mathbf{q}_K}(\mathbf{y}^e_{\mathrm{B}_K}) \end{bmatrix} \tag{29}$$

Proof:
From Eq. 26 we have:

$$\hat{\mathbf{x}}(\mathbf{y}^e) = (B^T B)^{-1} \sum_{k}^{K} B_k^T S_{\mathbf{q}_k}(\mathbf{y}^e_{\mathrm{B}_k})\mathbf{p}_k = \tilde{L}\mathbf{p} \tag{30}$$

The rest of the proof is similar to the proof given in Claim 1, by substituting matrix L with matrix $\tilde{L}$ $\square$.

The above claim gives the optimal K SFs to be applied to K different wavelet bands. Note, that in the un-decimated transform cases the term $B_k^T S_{\mathbf{q}_k}(\mathbf{y}^e_{\mathrm{B}_k})$ can be calculated efficiently by applying a 2D convolution to each of the slicing images composing the columns of $S_{\mathbf{q}_k}(\mathbf{y}^e_{\mathrm{B}_k})$.

6 Implementation Considerations

The theoretical structure of the suggested approach was introduced above. However, several computational issues are critical to the implementation of the approach and 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 $\tilde{L}^T\tilde{L}$ in Eq. 28 then becomes singular or ill-posed. In order to resolve this problem one must incorporate a regularization term in the minimization scheme. Referring to Eq. 27 we add a regularization term as follows:

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} \left\{ \|\hat{\mathbf{x}}^e - \mathbf{x}^e\|^2 + \sqrt{\lambda}\|\mathbf{p} - \mathbf{q}\|^2 \right\}$$

The regularization term deviates the solution of $\mathbf{p}$ towards $\mathbf{q}$ in particular in $\mathbf{p}$ entries where limited sampled data is available. As a result, the mapping function will be close to the identity in such $\mathbf{p}$ values. The constant parameter λ controls the influence strength of the regularization term. 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. It can be easily verified that the final solution of this system gives:

$$\hat{\mathbf{p}} = (\tilde{L}^T\tilde{L} + \lambda I)^{-1}(\tilde{L}^T\mathbf{x}^e + \lambda\mathbf{q}) \tag{31}$$

where I denotes a $K(M+1) \times K(M+1)$ identity matrix.

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 an additional term to the SLT definition (Eq. 15):

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

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

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

Inserting this term into Eq. 30 gives:

$$\hat{\mathbf{x}}^e = \tilde{L}\mathbf{p} + \tilde{\mathbf{h}}^e \quad \text{where} \quad \tilde{\mathbf{h}}^e = (B^TB)^{-1}\sum_k B_k^T\mathbf{h}^e_k \tag{32}$$

This updates the final solution which now gives:

$$\hat{\mathbf{p}} = (\tilde{L}^T\tilde{L} + \lambda I)^{-1}(\tilde{L}^T(\mathbf{x}^e - \tilde{\mathbf{h}}^e) + \lambda\mathbf{q}) \tag{33}$$

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 mapping functions. 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 incorporating all image equations together into a single equation and proceeding as above. If there are t example images denoted $\mathbf{x}_1^e \cdots \mathbf{x}_t^e$, Eq. 32 is extended to:

$$\begin{bmatrix} \tilde{L}_1 \\ \vdots \\ \tilde{L}_t \end{bmatrix} \mathbf{p} + \begin{bmatrix} \tilde{\mathbf{h}}_1^e \\ \vdots \\ \tilde{\mathbf{h}}_t^e \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{x}}_1^e \\ \vdots \\ \hat{\mathbf{x}}_t^e \end{bmatrix}$$

where $\tilde{L}_i$ and $\tilde{\mathbf{h}}_i^e$ are calculated as defined above for a single image.

In practice, however, a direct implementation of the above equation might be prohibited due to memory limitations. Note, that the dimensions of matrix $\tilde{L}$ in Eq. 29 is $N \times K(M+1)$ where N is the number of pixels in the example image. Fortunately, there is no need to construct the full equation in order to solve for $\mathbf{p}$. The solution $\hat{\mathbf{p}}$ minimizing the MSE cost function:

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} \sum_{i}^{t} \|\hat{\mathbf{x}}(\mathbf{y}_i^e) - \mathbf{x}_i^e\|^2$$

gives rise to the following solution which replaces Eq. 33 above:

$$\hat{\mathbf{p}} = \left(\sum_{i}^{t}(\tilde{L}_i^T\tilde{L}_i) + \lambda I\right)^{-1}\left(\sum_{i}^{t}\tilde{L}_i^T(\mathbf{x}_i^e - \tilde{\mathbf{h}}_i^e) + \lambda\mathbf{q}\right)$$

Note that the dimensions of $\tilde{L}_i^T\tilde{L}_i$ is $K(M+1)\times K(M+1)$ where commonly $K(M+1) << N$. Therefore, implementing the above solution needs only a memory capacity on the order of (KM^2) which is independent of the number of images. This scheme can be implemented also to a single image if its size is too large. In such a case, the image is split into several sub-images each of which 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 gives rise to symmetric SFs 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 $\mathbf{x} = S_{\mathbf{q}}(\mathbf{x})\mathbf{q} + \mathbf{r}$ is still correct if the definition of $S_{\mathbf{q}}(\mathbf{x})$ is modified as follows:

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

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 Slicing Transform. The greater the number of bins used the more flexibility we gain for the generated SFs (although at the expense of computational burden). It was experimentally demonstrated that relatively few quantization bins (15 bins in our experiment settings) produce superior results which are very close to the asymptotic quality (see results in Section 7). Additionally, since small wavelet values are much more probable that higher values, it is preferable to implemented a non-uniform 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 the mapping domain where a unit interval $t \in [0,1]$ is transformed to the mapping domain $s \in [a,b]$ through a polynomial function: $s = t^\beta \cdot (b-a) + a$, for $\beta > 1$.

7 Results

In order to demonstrate the advantages of the proposed approach and to indicate the source of improvements, we compare the de-noising results using three different schemes:

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

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

  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:

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

  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 [8, 5, 31] with the exception that the SFs are extracted in a discriminative manner rather than a descriptive one.

- **Method 2** (spatial domain - independent bands): A set of SFs is optimized in the spatial domain. The objective term for this method reads:

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

  This minimization gives rise to the following solution:

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

  Note, that the objective criterion is expressed in the spatial domain, yet, the SFs are evaluated for each band independently. Thus, although within-band dependencies are exploited through the backward transform, this method still ignores inter-band dependencies.

- **Method 3** (spatial domain - joint bands): The scheme suggested in this paper (Sec. 5) where the objective goal is expressed in the spatial domain:

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

  and the solution is given in Eq. 28. In this scheme the SFs are evaluated simultaneously while taking into account inter-band as well as intra-band dependencies.

In all the experiments described below we used the undecimated windowed DCT as the image transform. Since the undecimated DCT is a tight frame, the term $(B^T B)^{-1}$ in Eq. 29 is simplified: $(B^T B)^{-1} = (\sum_i B_i^T B_i)^{-1} = \frac{1}{K} I$, enabling efficient implementation. Note, that due to the undecimated form each wavelet band can be calculated using a single 2D 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 kernel.

In the following experiments, unless mentioned otherwise, the setting parameters were defined as follows: (1) Test images were taken from Figure 5. (2) Training was performed on the top-right image of Figure 6. (3) Transform basis was the undecimated 8×8 DCT. (4) The noise consists of additive Gaussian noise with s.t.d. of 20 gray levels.

Figures 7-10 display some of the SFs obtained for 8×8 DCT basis, using the three methods described above. Figure 7 shows the SFs resulting from the 1st method. It can be seen that these SFs resemble the traditional SFs where the mapping functions are monotonic non-decreasing (see e.g. [29]). SFs shown in Figures 8 and 9 present the results of the the 2nd and the 3rd method respectively. In contrast to the traditional methods, here, the

produced SFs do not necessarily retain monotonicity and the mapping functions may descend below zero or rise above the identity line. In particular, the SFs presented in Figure 9 have portions in which positive coefficients are mapped to negative values and portions in which negative coefficients are mapped to positive values, requiring regions having a negative slope. Since this behavior does not appear in Methods 1 and 2, it might be concluded that this phenomena is due to the band dependencies that in taken into account in Method 3. For comparison purpose, Figure 10 shows side-by-side some of the produced SFs using methods 1-3.

We tested the obtained SFs on several images shown in Figure 5. These images are commonly used as test cases for denoising algorithms[1]. Figure 11 compares the resulting psnr for each one of the described methods. The figure is composed of 6 clusters of bars, each of which compares the denoising results of a particular image. Each bar presents the denoising results averaged over 10 realizations of noise with s.t.d. of 20 gray levels. The results demonstrate the improvement of the 2nd method over the 1st method, and the superiority of the 3rd method over the other two. Note, that the traditional approaches which optimize the SFs in the transform domain are comparable with the 1st method. It can be seen that most of the improvement is achieved due to formulating the objective in the spatial domain (Method 2). Additional improvement, although less significant, is achieved when incorporating the band dependencies into the solution (Method 3). Examples of denoised images after applying Method 3 to the test images are shown in Figure 12.

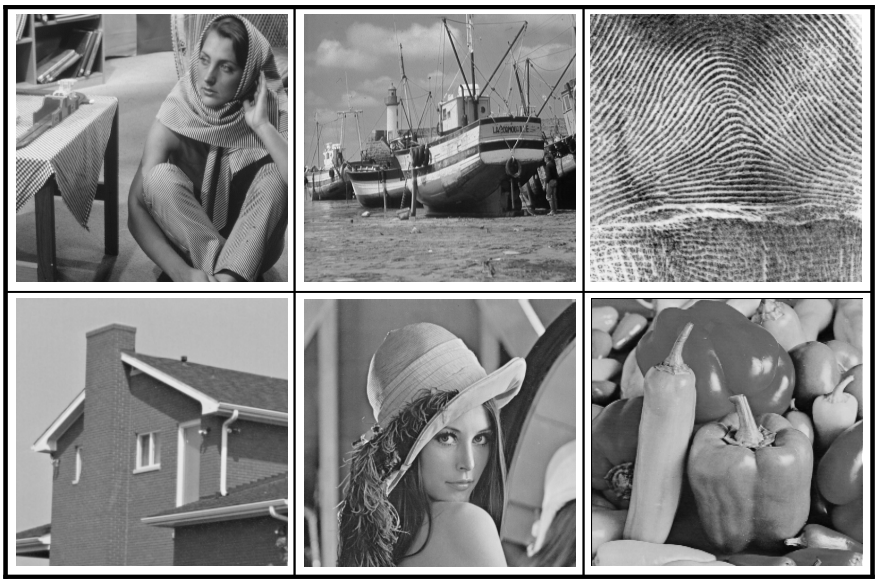

Figure 5: The images on which the denoising schemes were tested. From left to right top down: BARBARA, BOAT, FINGERPRINT, HOUSE, LENA, PEPPERS.

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

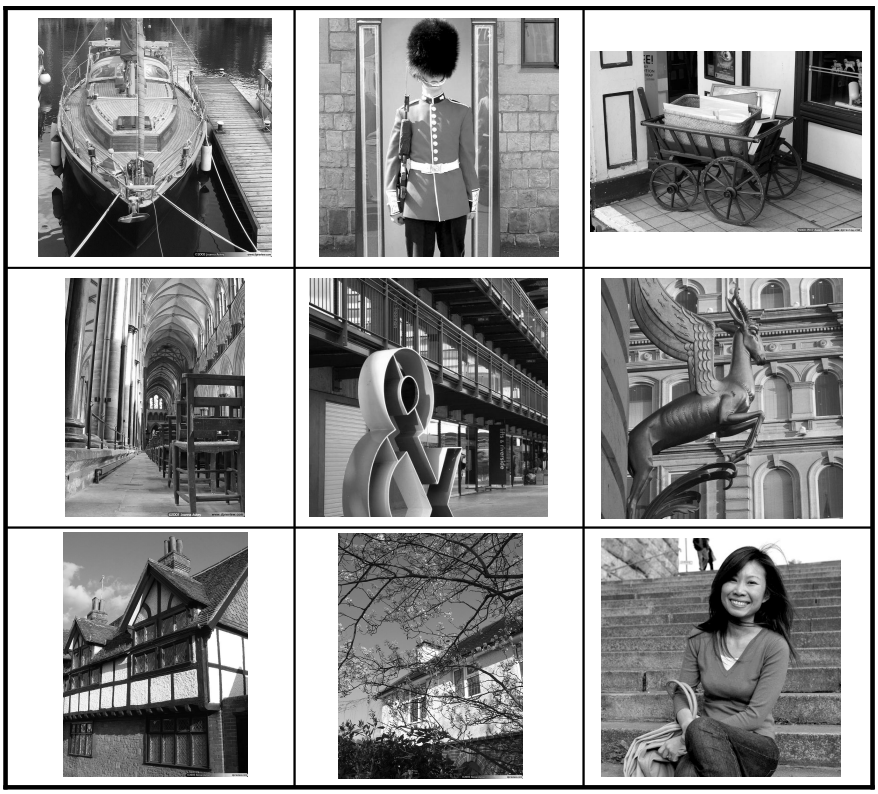

Figure 6: The images on which the SFs were trained.

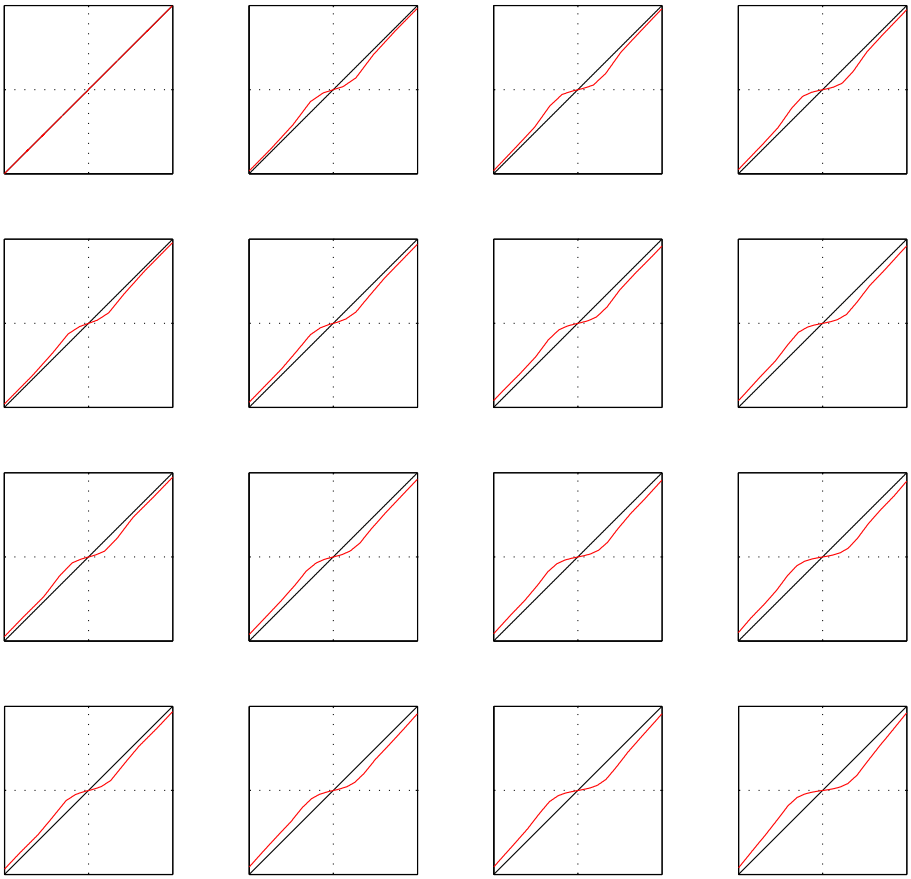

Figure 7: SFs produced using Method 1. The filters used were 8×8 DCT. The SFs shown are for the DCT bands whose indices are: [1..4] $\times$ [1..4] (left to right - top to bottom). SFs are shown in the range [-120,120] at each axis.

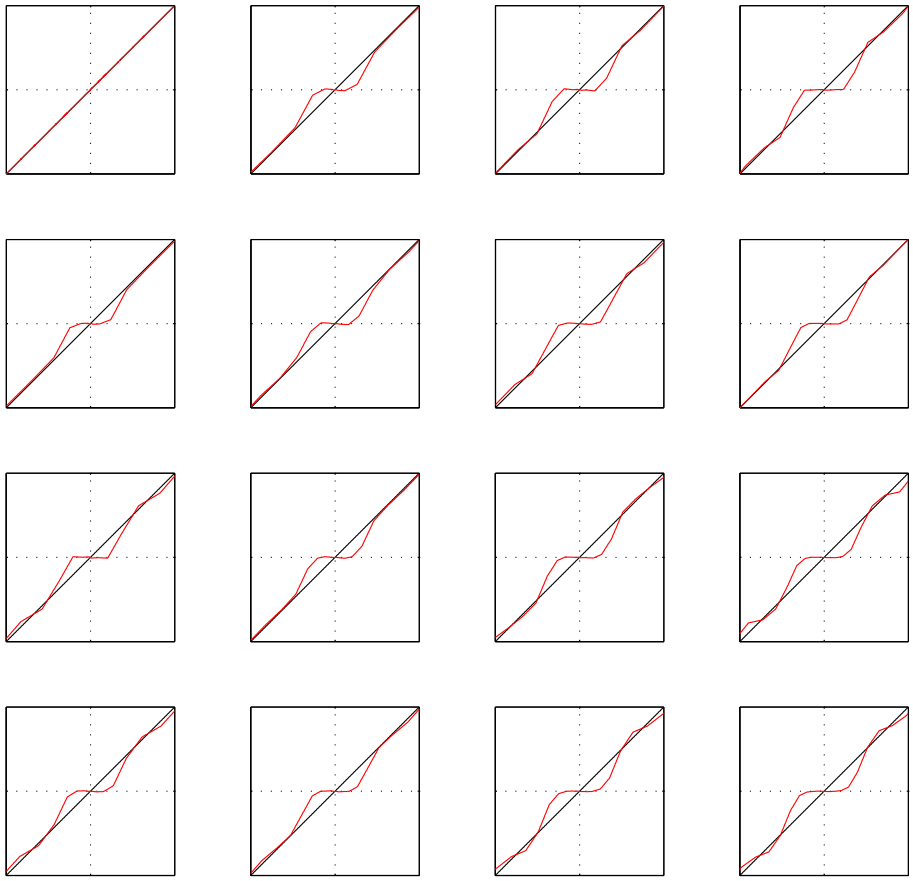

Figure 8: SFs produced using Method 2. The filters used were 8×8 DCT. The SFs shown are for the DCT bands whose indices are: [1..4] $\times$ [1..4] (left to right - top to bottom). SFs are shown in the range [-120,120] at each axis.

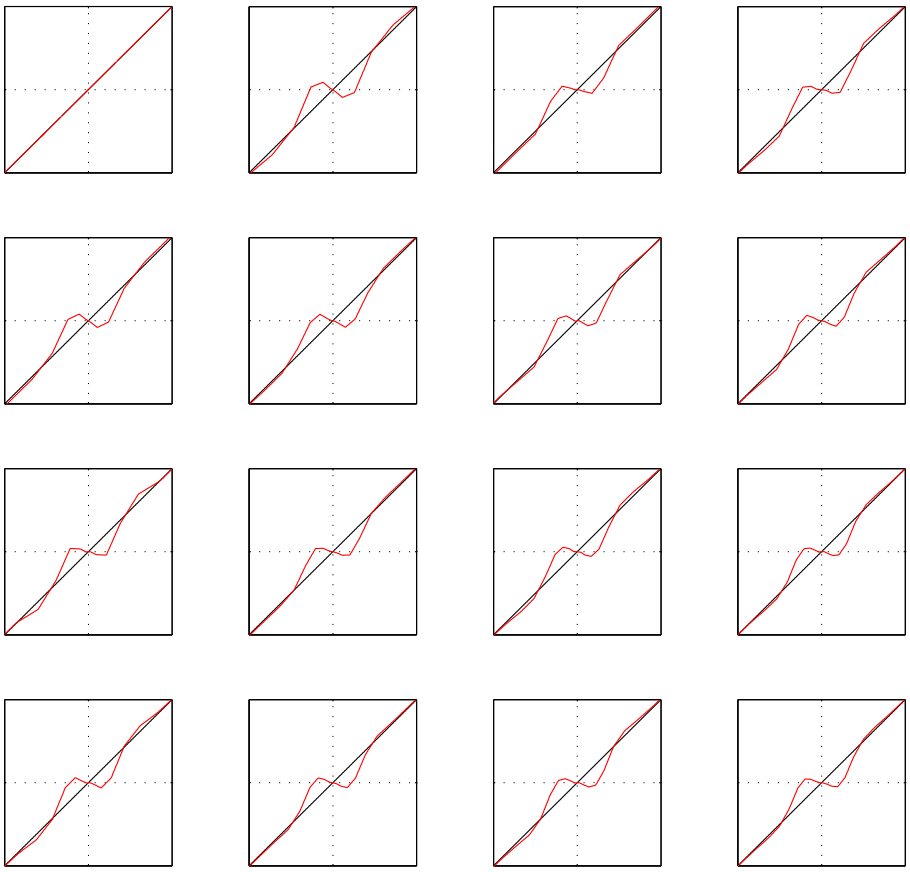

Figure 9: SFs produced using Method 3. The filters used were 8×8 DCT. The SFs shown are for the DCT bands whose indices are: [1..4] $\times$ [1..4] (left to right - top to bottom). SFs are shown in the range [-120,120] at each axis.

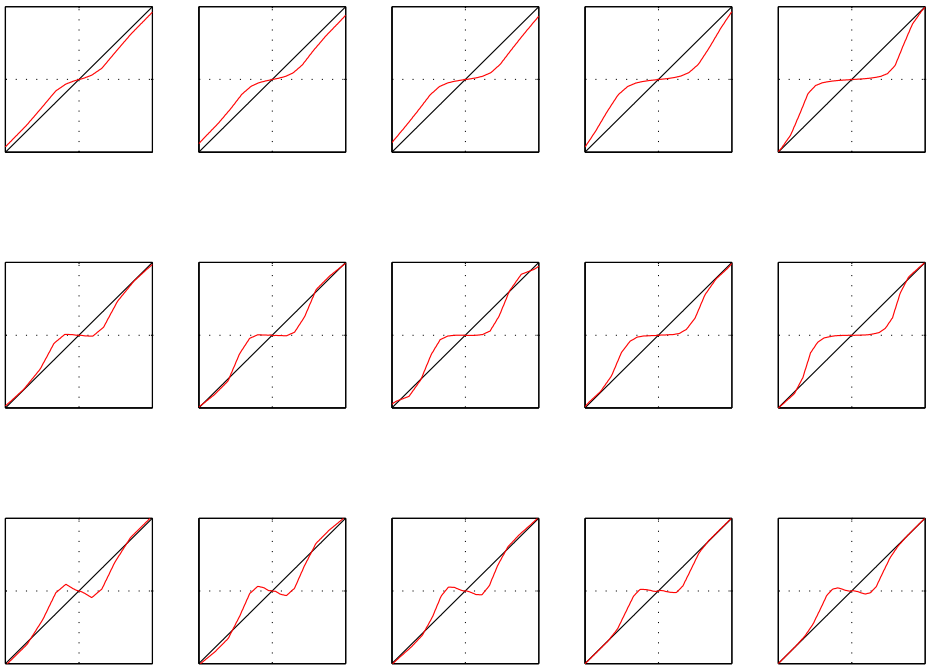

Figure 10: Comparison of the produced SFs using Method 1 (top row), Method 2 (middle row), and Method 3 (bottom row). SFs on each row correspond to band (i,i) of the 8×8 DCT basis, where $i = 2..6$ (left to right). Graph axes are shown in the range [-120,120].

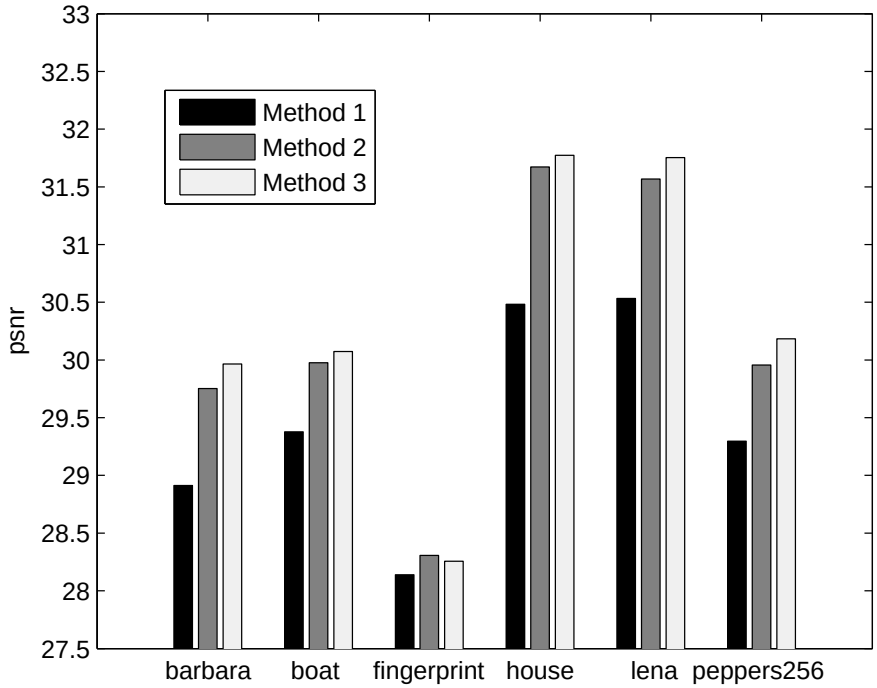

Figure 11: Psnr after applying the SFs produced by methods 1-3. Each bar is an average over 10 different noise realizations.

7.1 The noise variance

The influence of the noise variance on the obtained SFs can be seen in Figure 13. Similarly to the classical hard/soft thresholding SFs, the profiles of the produced SFs 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 variance. Thus, if a particular SF $\mathcal{M}_{\sigma_0}\{v\}$ was obtained for noise variance σ_0, the same SF for noise variance σ is expected to be:

$$\mathcal{M}_{\sigma}\{v\} = s\mathcal{M}_{\sigma_0}\{\frac{v}{s}\} \qquad \text{where} \quad s = \frac{\sigma}{\sigma_0} \tag{34}$$

Representative examples of this behavior are demonstrated in Figure 14 where the sum of squared differences (SSD) between the SFs obtained for noise level 20 s.t.d. and other scaled SFs, obtained for different noise levels, are shown in colored diagrams. The diagrams show the log SSD as a function of x-scaling (horizontal axis) and y-scaling (vertical axis). It is demonstrated that uniform scalings produces the best matches between the SF pairs. The amount of best uniform scalings for various noise levels are shown in Figure 15. The graph shows a clear linear dependency between the measured optimal scaling and the relative noise

Figure 12: Some examples of de-noised images. The images in the top row were contaminated with white noise with s.t.d. of 20 gray-levels. The reconstructed images are shown on the bottom row.

variance, i.e. $\sigma/20$, which follows the expression given in Eq. 34. An example of two sets of SFs, 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 Figure 16. It is demonstrated that the two SFs coincide almost perfectly and are hard to distinguished.

7.2 The Training Images

The resemblance between the training images and the target noisy images plays a significant role in the denoising quality. The importance of this factor is demonstrated in Figure 11 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 Figure 6) 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 FINGERPRINT image rotated by 180°). The results are given in Figure 17. 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

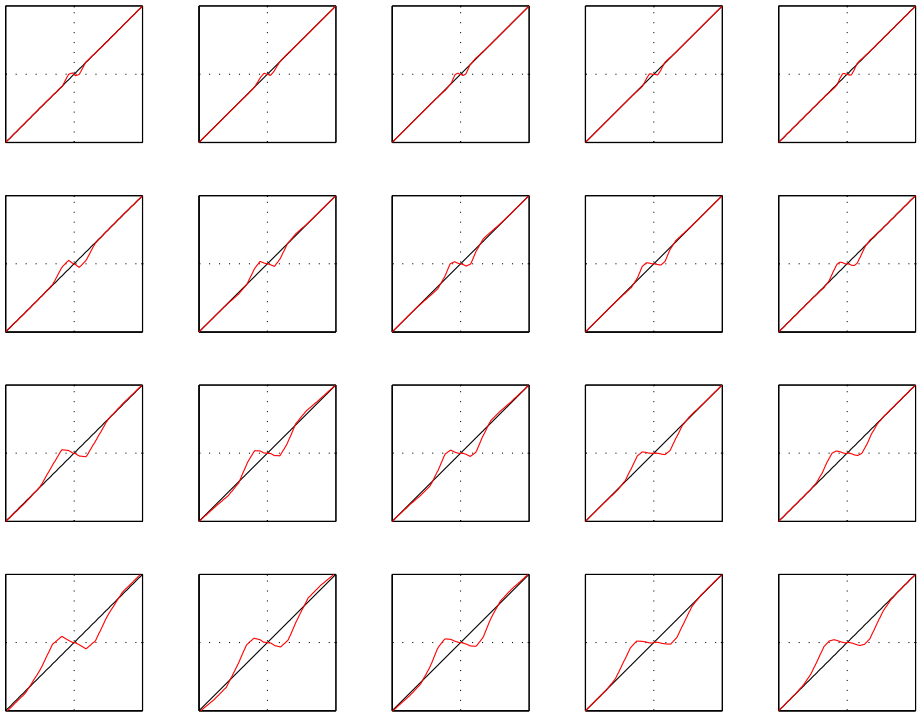

Figure 13: SFs produced using Method 3 for various noise levels. SFs on each row correspond to band (i,i) of the 8×8 DCT basis, where $i = 2..6$ (left to right). The noise levels were of 5, 10, 15, and 20 s.t.d. from top row to bottom row, respectively. Graph axes are shown in the range [-120,120].

db.

Although the choice of good training images poses a challenging problem, this difficulty comes with an important advantage as one can fit the produced SFs to the type of noisy images that are to be denoised. A general approach evaluating the level of resemblance between a target image and its training set is outside the scope of this paper. In our experiments we used a fixed set of natural images, however, it is reasonable that improvement in the results may be achieved if the training images are selected in an adaptive manner. Figure 18 presents a set of resulting psnr using 9 different SFs each of which was trained on a different natural image taken from the set shown in Figure 6. In this experiment the choice of the trained natural image influenced the resulting psnr by up to 0.2 db.

Another factor that might influence the quality of the solution is the size of the training image or alternatively the number of images the SFs are trained from. The next test tries to

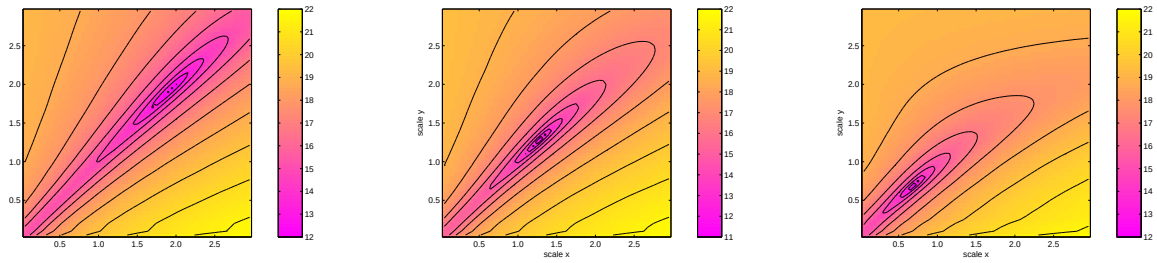

Figure 14: The accumulated differences (measured in log SSD) between the SFs obtained for noise level 20 s.t.d. and other scaled SFs obtained for noise levels 10 s.t.d. (left), 15 s.t.d. (middle), and 25 s.t.d. (right). Each diagrams present the SSD as a function of x-scaling (horizontal axis) and y-scaling (vertical axis). It is demonstrated that a uniform scaling produces the best matching.

answer this question experimentally. In this experiment the training images are a set seven natural images taken from Figure 6. Seven sets of SFs were generated each of which was learnt from a different number of example images. The denoising results are shown in Figure 19. The bars for each image are associated with a different number of training images ranging from 1 image (left most) to 7 images (right most). These results show that increasing the number of training images does not necessarily improve denoising quality, and that a single training image is sufficient to capture the image statistics. Nevertheless, a theoretical bound for the generalization power of the learnt SFs with respect to new images does not exist and should be further investigated. This issue is closely related to the generalization power in classification techniques where the goal is to design optimal classifiers whose generalization capabilities are assured [34].

7.3 The Transform Used

Previous studies demonstrated the benefit of using particular transforms, such as steerable pyramids, curvelets, and contourlet [31, 32, 7] as being more appropriate for modeling natural images. The scheme presented in this paper is general, and can work with any given transform or any set of filters (such as the bilateral filters [33]). In all our experiments we used the undecimated DCT transform with various window sizes. As will be shown below, the obtained results demonstrate quality comparable with the state-of-the-art methods. It is expected that further improvement can be achieved if other, more appropriate, transforms will be used. Figure 20 presents the denoising results using the undecimated DCT transform with various window size. 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.

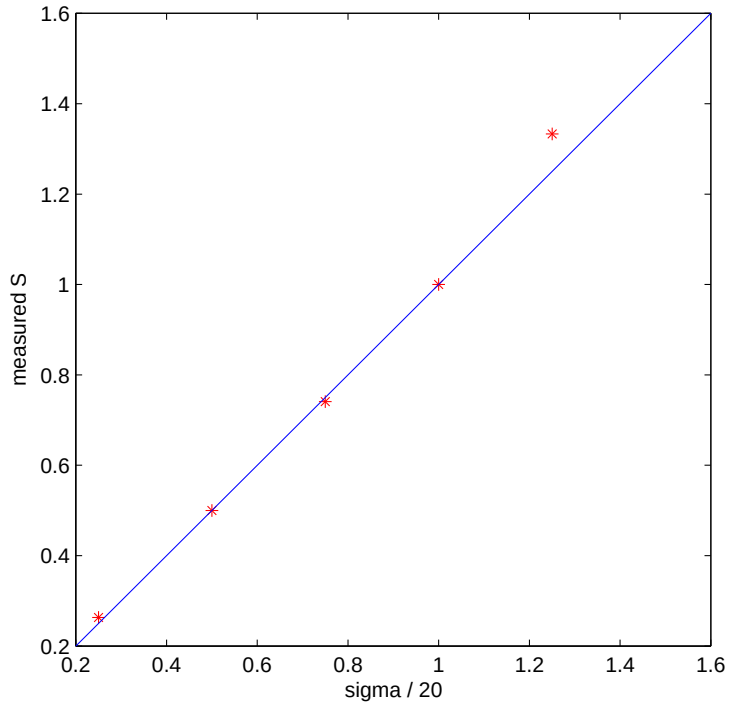

Figure 15: The optimal (uniform) scaling producing the best matches between the 20 s.t.d. SFs and other noise level SFs. The graph shows a linear dependency between the measured optimal scaling and the relative noise variance, i.e. $\sigma/20$.

7.4 Number of Quantization Bins

Figure 21 shows the resulting psnr v.s. the number of quantization bins used. It is shown that around 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 SFs. It also strengthens the rationale behind modeling the SFs 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 that higher values, we implemented a non-uniform quantization bins as elaborated in the previous section.

7.5 Comparison with Other Methods

The proposed approach was tested on the images presented in Figure 5 which were contaminated with white noise under various noise levels. The resulting psnr are shown in Table 1. The transform used in this table was the undecimated 9x9 DCT. Although the transform used is not optimal for natural images and the training image was chosen arbitrarily

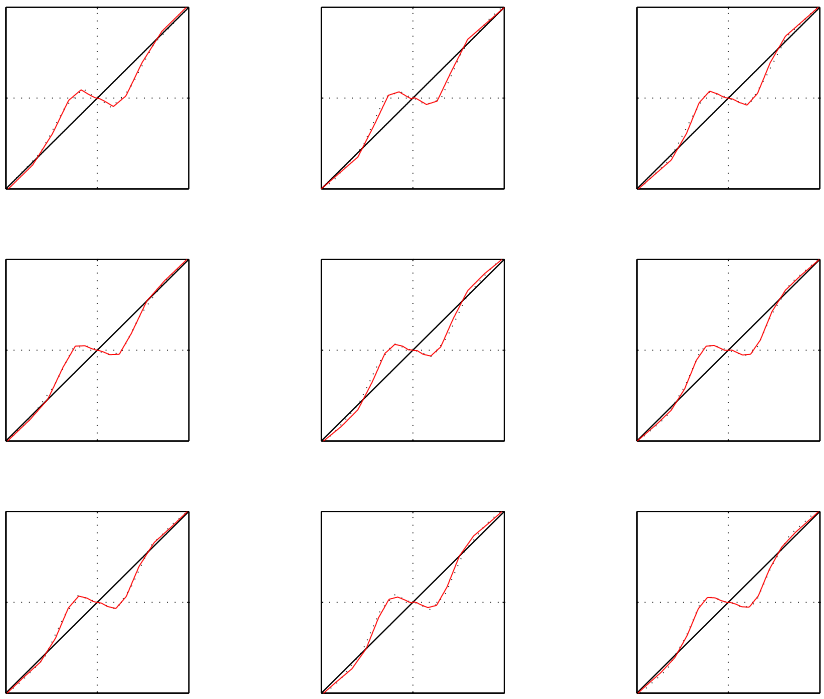

Figure 16: A comparison between the SFs produced for 20 s.t.d. (solid red) and 10 s.t.d. scaled by 2 (dotted black). The SFs shown are for DCT bands [2..4] $\times$ [2..4]. The graph axes are shown in the range $[-120, 120]$. The two graphs coincide almost perfectly and hard to distinguished.

(top-right image of Figure 6) 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 Figure 22 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 where in more severe noise cases (15 s.t.d. and above) the BLS-GSM demonstrates marginally better performance. Comparison results averaged over all test images are shown in Figure 23.

7.6 Other Reconstruction Problems

The approach described in this paper is presented in the context of image denoising where the contaminated noise is assumed white. However, since the approach does not require

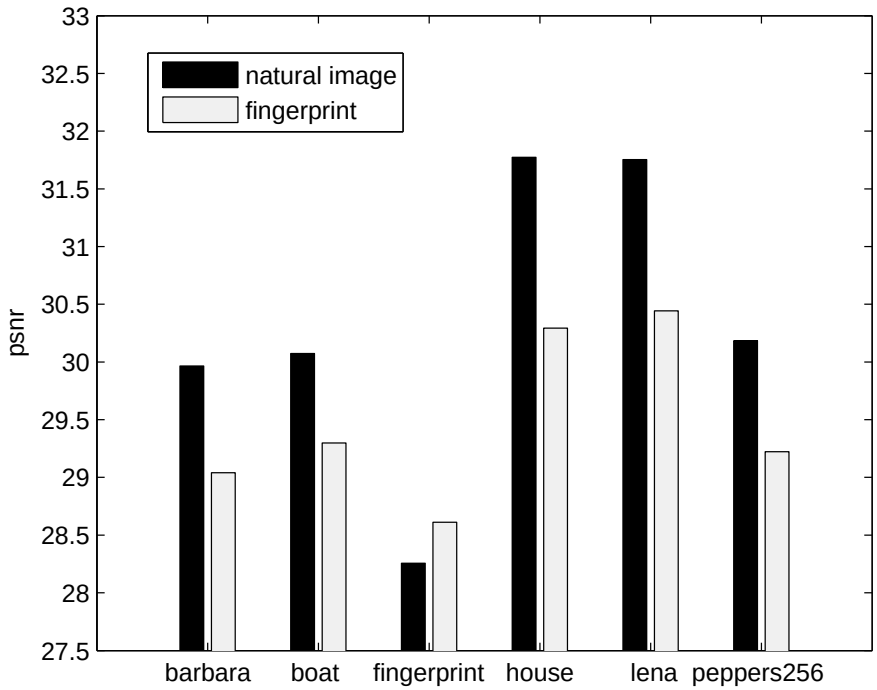

Figure 17: Comparison results for denoising images using method 3 where the training images were: FINGERPRINT rotated by 180° (gray bars) and a natural image (black bars).

any modeling of the image statistics or of the noise, it can be seamlessly applied in other reconstruction problems and with different types of noise characteristics. As long as the reconstruction process involves applying marginal look-up-tables in the transform domain, optimal SFs can be obtained. One only needs to provide pre-degradation and post-degradation images. This section presents some examples of applying the suggested approach to other reconstruction problems, namely: removing JPEG artifacts, and image de-blurring. These examples are given in order demonstrate the concept with no comparative study.

In the first experiment we attempted to de-blur images using a set of look-up tables (LUT) 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 Figure 24. The full set of LUTs were applied to a blurred version of BARBARA (same blurring parameters). A close-up view of the de-blurred BARBARA is given in Figure 25. The resulting image demonstrates very good 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

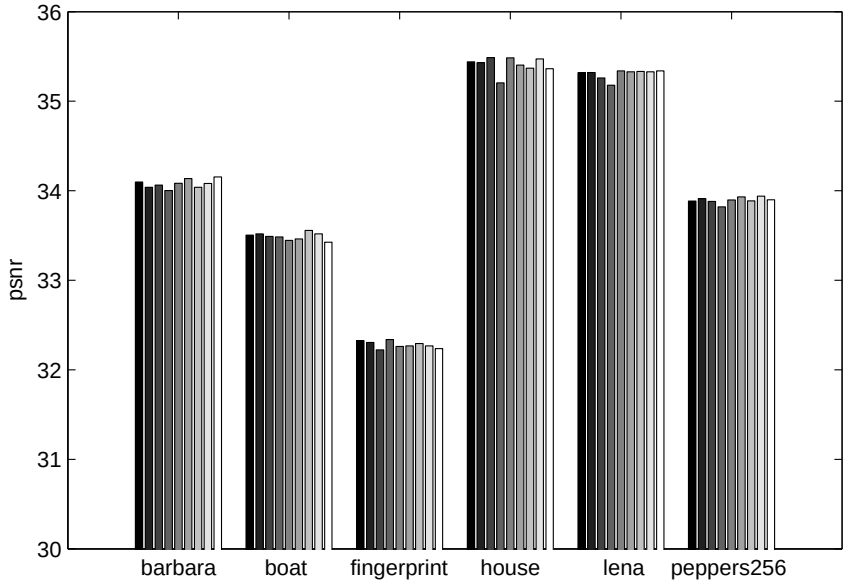

Figure 18: The psnr of denoised images using SFs that were trained on various natural images shown in Figure 6. The transform used was DCT 8×8. The contaminated noise was white noise with 10 s.t.d. gray-levels.

LENA. The "noise" was generated by JPEG-compression with quality parameter set to 30%. A partial set of the LUTs are shown in Figure 26. The JPEG artifacts of the compressed image are presented in Figures 27-left and in a close-up view in Figure 28-top. The artifact removal after applying the learnt LUTs are shown in Figure 27-right and Figure 28-bottom. The quality of the reconstruction is self-evident. It is interesting to mention the resemblance of the proposed approach to that of Nosratinia [21]. 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 to designed a new set of LUTs that are optimized to produce the best results.

8 Discussion and Conclusions

This paper suggests a new and simple scheme for Wavelet denoising relying on a discriminative concept. 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. [26, 13, 30]). Using the suggested scheme, however, we

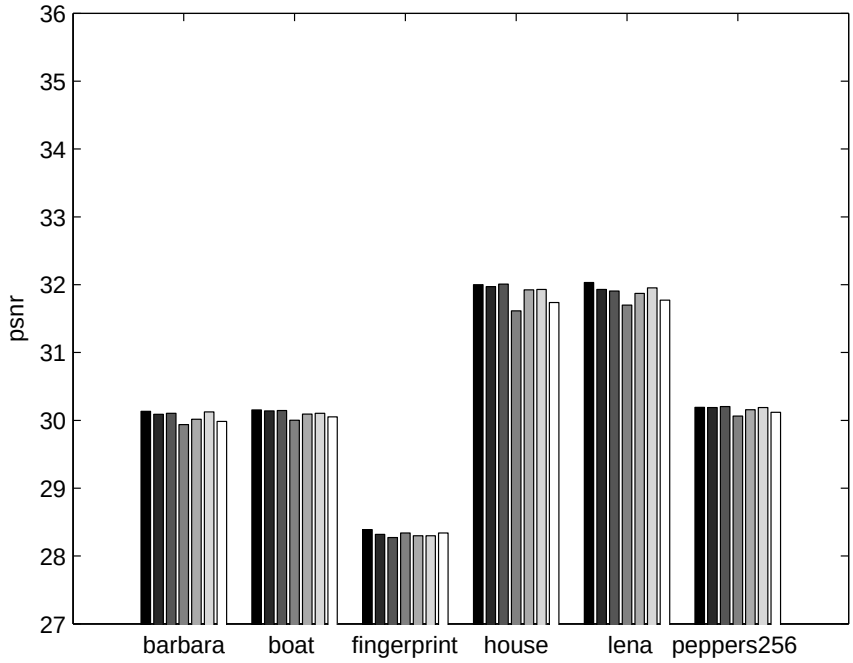

Figure 19: The psnr of denoised images v.s. the number of training images SFs were trained from. Each group of bars shows the psnr arising from different sized training sets, ranging from 1 to 7 (left to right).

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. Again, in our approach we assume only the existence of a noise model and perform similarly whether the true noise is simple white Gaussian or more complex (e.g. JPEG noise). Thus, the discriminative approach enables us to apply the proposed scheme to other degradation processes seamlessly. As long as the restoration process relies on marginal rectification of transform coefficients the suggested scheme produces the optimal solution. This optimality is achieved with respect to several aspects:

- An optimal set of scalar SFs are generated for over-complete transforms taking into account intra-band and inter-band dependencies. As far as we know, previous scalar SF 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

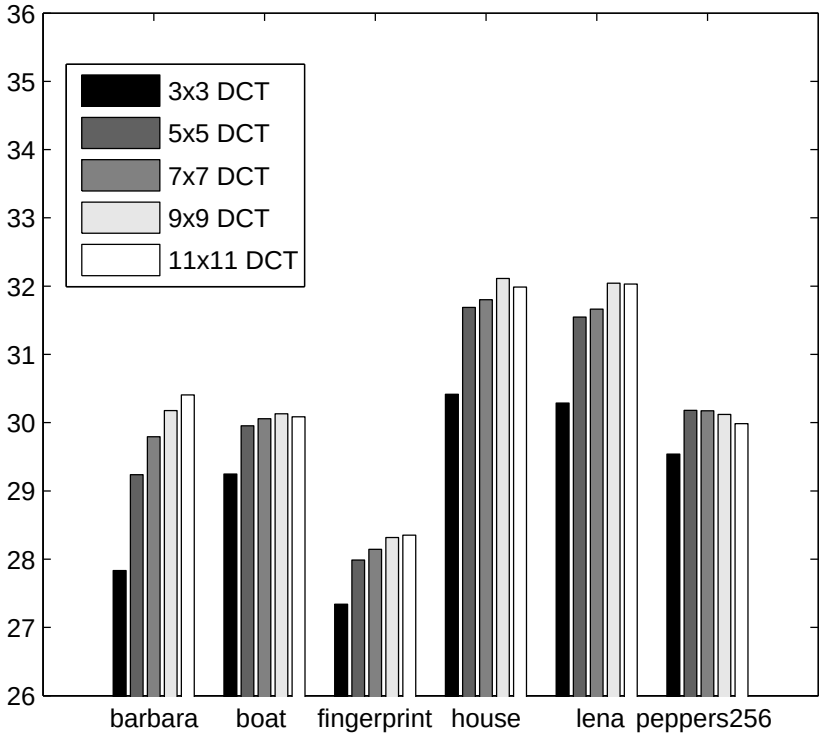

Figure 20: The resulting psnr v.s. DCT window size.

hurdle in the descriptive approach, in the proposed discriminative model the restriction to the spatial domain posed only a computational burden.

- The proposed scheme expresses the optimal objective with respect to the MSE criterion rather than the MAP which is the common criterion applied in descriptive approaches. Although parametric SFs optimizing the MSE criterion were previously proposed (e.g. [29]) these were developed in the transform domain, thus, do not guarantee optimality in the spatial domain.

One of the main contributions in this paper is the formulation of the Slicing Transform in which non-linear operations can be applied in a linear manner. We believe this formulation can be useful beyond the scope of denoising, and it can be further exploited in other problems such as as tone-correction, image metrics, color mapping, and more.

As emphasized above, the suggested scheme is based on marginal rectification of transform coefficients, namely the SFs 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 (on-line). This restriction can be relaxed by applying multivariate SFs, possibly

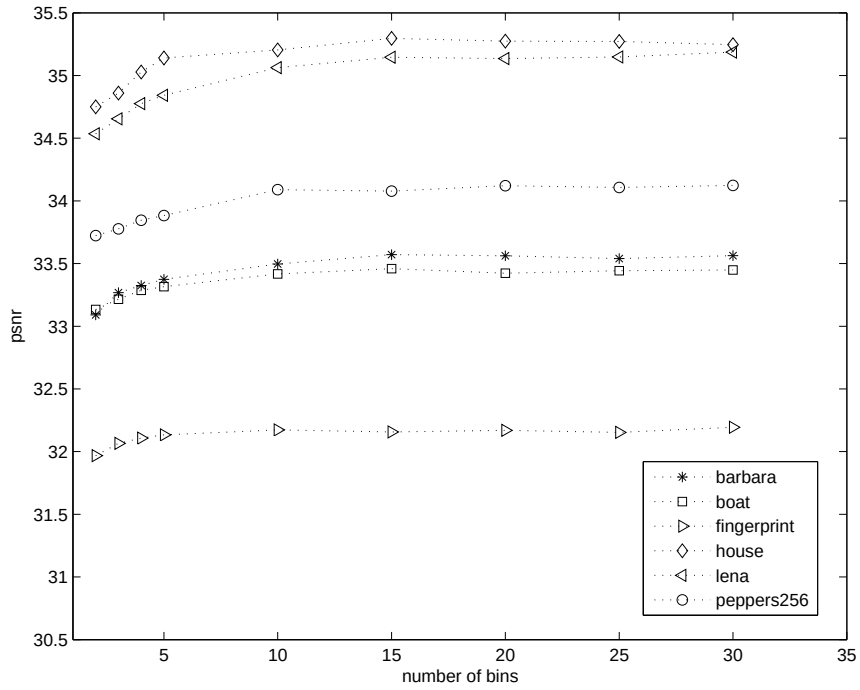

Figure 21: The resulting psnr v.s. the number of quantization bins used for the SFs. The transform used was the undecimated 5x5 DCT. The results are shown for various images with noise s.t.d.=10.

approximated 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 is 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 design a set of SFs that are selected adaptively according to the estimated local noise variance.

There are two important issues that were not dealt within 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 7.3). 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 transform) that produces the best results is still an open question.

Noise s.t.d.	BARB.	BOAT	FGRP.	HOUSE	LENA	PEPPERS256
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

Table 1: Resulting psnr for various noise levels. The transform used was the undecimated 8x8 DCT. The SFs were trained on the top-right image in Figure 6.

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 between the test and the training images. Thus, SFs for cartoon type images, for example, should be trained on cartoon examples and SFs trained on a particular texture should be applied to similarly textured images.

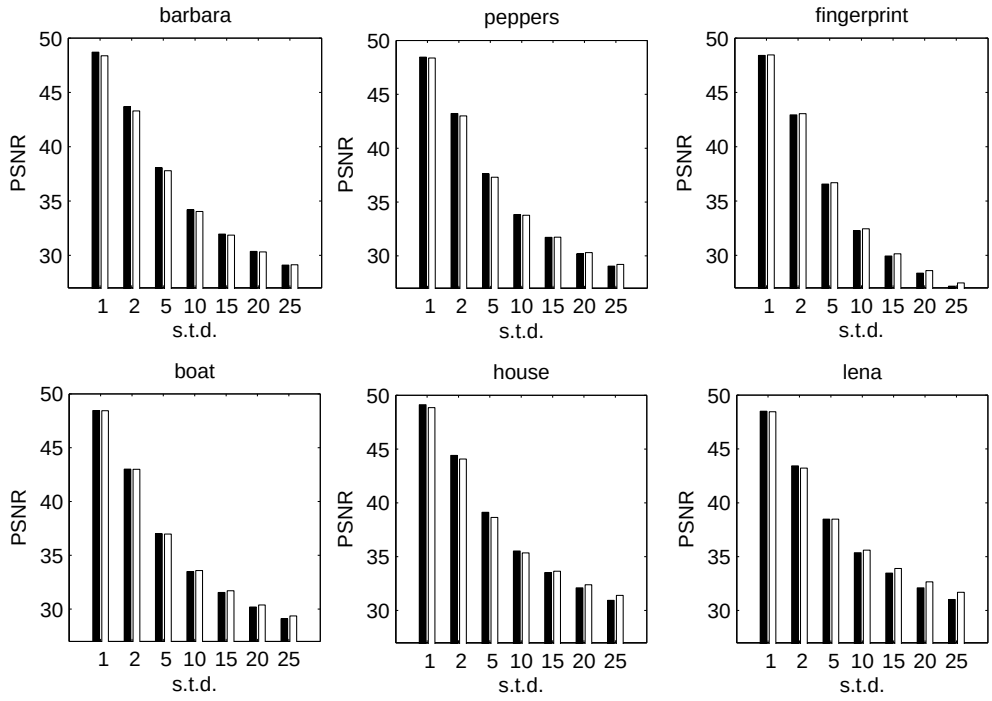

Figure 22: Comparison between the proposed method and the BLS-GSM method for various noise levels. Dark bars: The porposed method. White bars: The BLS-GSM method.

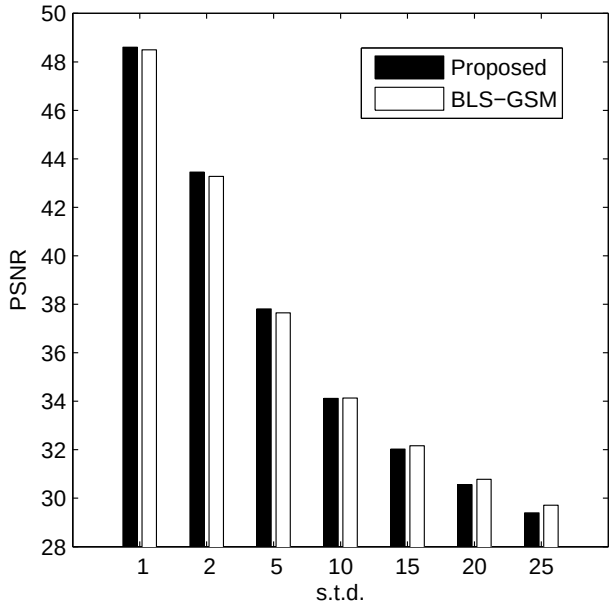

Figure 23: Comparison between the proposed method and the BLS-GSM method averaged over all test images.

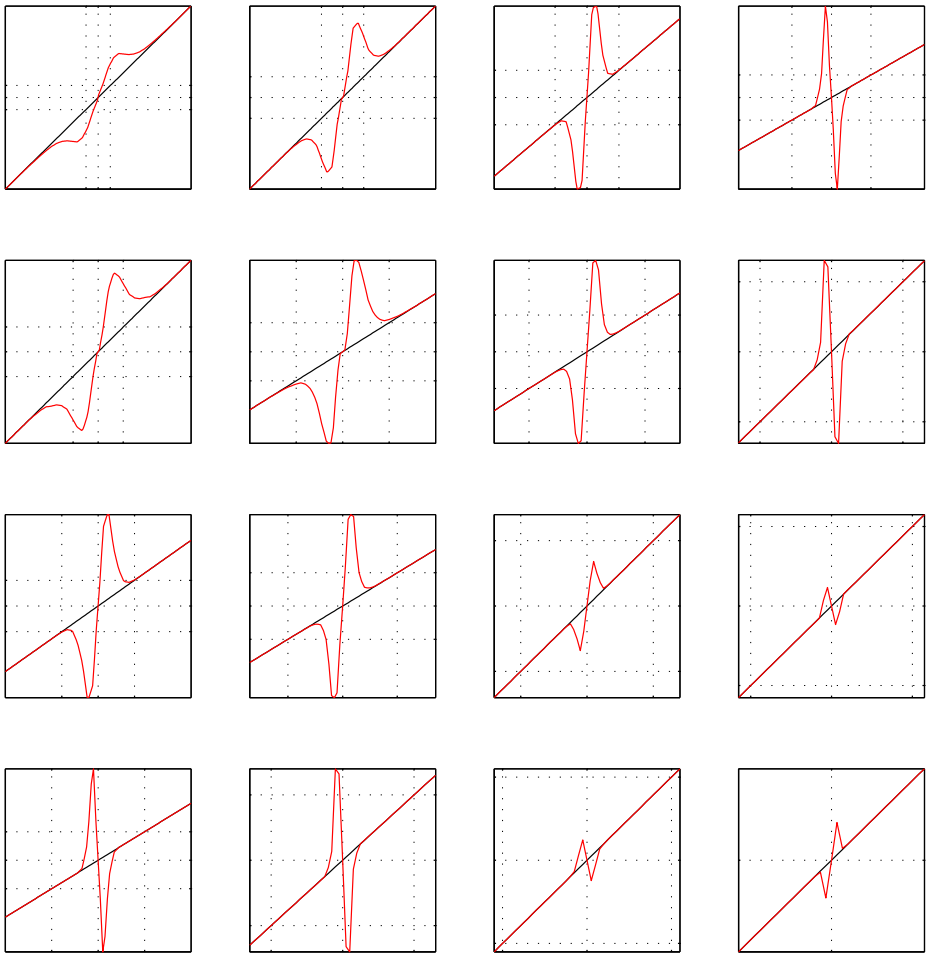

Figure 24: The SFs that were learnt to sharpen images blurred with a 5-tap Gaussain kernel. The filters used were 8x8 DCT. The SFs shown are for the DCT bands whose indices are: $[3..6] \times [3..6]$ (left to right $\times$ top to bottom). The scaling factor of each graph can be indicated by the dotted lines, plotted at values $[-20\ 0\ 20]$ for each axis.

Figure 25: Left: Blurred BARBARA after applying a 5-tap Gaussian blur. Right: Sharpened BARBARA using LUTs that were trained on blurred LENA using 8x8 over-complete DCT.

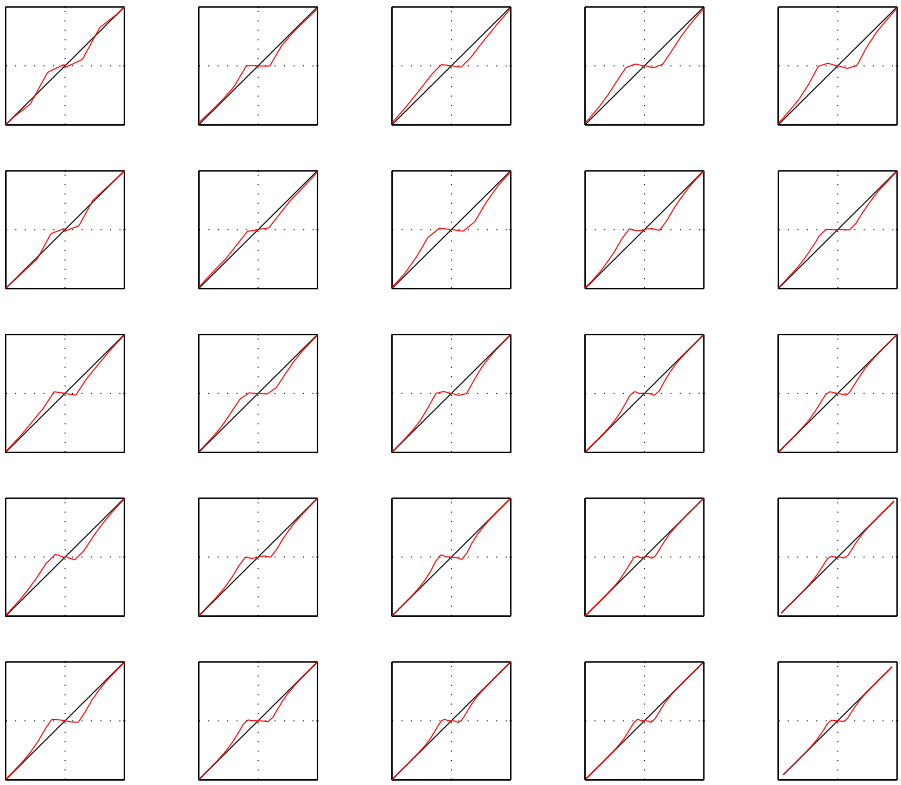

Figure 26: The LUTs that were learnt from JPEG-compressed BARBARA with quality=30. The filters used were DCT 8x8. The SFs shown are for the DCT bands whose indices are $[4..8] \times [4..8]$ (left to right $\times$ top to bottom). Graph axes are shown in the range $[-60, 60]$.

Figure 27: Left: JPEG artifacts after compressing LENA with JPEG quality=30. Right: Artifact removal using LUTs that were trained from JPEG-compressed BARBARA using the 8x8 DCT.

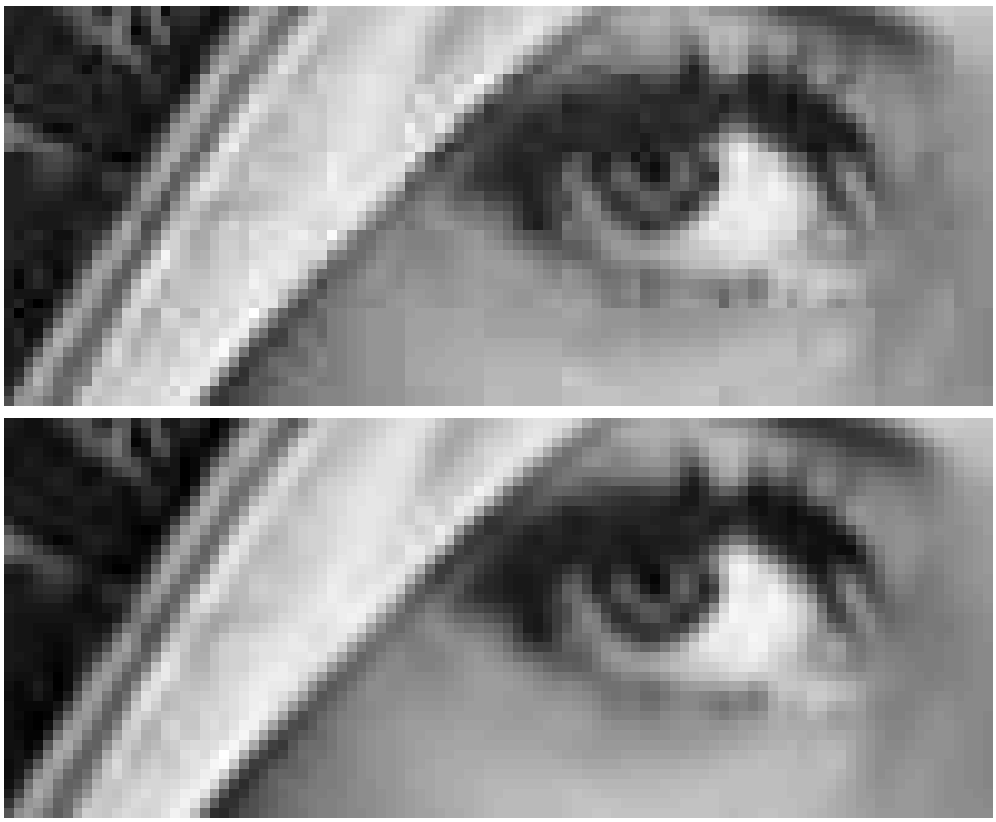

Figure 28: Top: A zoom-in of LENA JPEG artifacts. Bottom: A zoom-in of the artifact removal using the proposed method.

Appendix:

A Minimization Domain

Consider finding the LS optimal SFs for the unitary case. Namely, finding $\vec{\mathcal{M}}_{\mathrm{w}}$ that minimizes

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

where $\hat{\mathbf{x}}(\mathbf{y}) = W^T\vec{\mathcal{M}}_{\mathrm{w}}\{W\mathbf{y}\}$. Whenever W is unitary this minimization can be expressed equivalently in the transform domain:

$$\begin{aligned}
\varepsilon &= E_{\mathbf{x}|\mathbf{y}}\{\|W^T\vec{\mathcal{M}}_{\mathrm{w}}\{W\mathbf{y}\} - \mathbf{x}\|^2\} \\
&= E_{\mathbf{x}|\mathbf{y}}\{(W^T\vec{\mathcal{M}}_{\mathrm{w}}\{W\mathbf{y}\} - \mathbf{x})^T(W^T\vec{\mathcal{M}}_{\mathrm{w}}\{W\mathbf{y}\} - \mathbf{x})\} \\
&= E_{\mathbf{x}|\mathbf{y}}\{(W^T\vec{\mathcal{M}}_{\mathrm{w}}\{W\mathbf{y}\} - \mathbf{x})^T W^T W(W^T\vec{\mathcal{M}}_{\mathrm{w}}\{W\mathbf{y}\} - \mathbf{x})\} \\
&= E_{\mathbf{x}|\mathbf{y}}\{(WW^T\vec{\mathcal{M}}_{\mathrm{w}}\{W\mathbf{y}\} - W\mathbf{x})^T(WW^T\vec{\mathcal{M}}_{\mathrm{w}}\{W\mathbf{y}\} - W\mathbf{x})\} \\
&=^* E_{\mathbf{x}|\mathbf{y}}\{(\vec{\mathcal{M}}_{\mathrm{w}}\{W\mathbf{y}\} - W\mathbf{x})^T(\vec{\mathcal{M}}_{\mathrm{w}}\{W\mathbf{y}\} - W\mathbf{x})\} \\
&= E_{\mathbf{x}|\mathbf{y}}\{\|\vec{\mathcal{M}}_{\mathrm{w}}\{\mathbf{y}_{\mathrm{w}}\} - \mathbf{x}_{\mathrm{w}}\|^2\}
\end{aligned}$$

On the other hand, for an overcomplete transform matrix W, we have that $\frac{1}{N}WW^T \neq I$, which implies that the derivation above fails at $*$, and the minimization goal cannot be expressed equivalently in the transform domain.

B References

[1] E. J. Candes. Harmonic analysis of neural networks. *Applied and Computational Harmonic Analysis*, 6:197–218, 1999.

[2] P. Carré and D. Helbert. Ridgelet decomposition: Discrete implementation and color denoising. In *Wavelet Applications in Industrial Processing III*, Boston, Massachusetts, USA, October 2005. SPIE.

[3] S. Chang, B. Yu, and M. Vetterli. Spatially adaptive wavelet thresholding with context modeling for image denoising. *IEEE Trans. Image Processing*, 9(9):1522–1531, September 2000.

[4] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomoc decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.

[5] R. R. Coifman and D. L. Donoho. Translation invariant de-noising. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, pages 125–150. Springer-Verlag, 1995.

[6] M. Crouse, R. Nowak, and R. Baraniuk. Wavelet-based statistical signal processing using hidden markov models. *IEEE Transactions on Signal Processing*, 46:886–902, 1998.

[7] M. N. Do and M. Vetterli. The contourlet transform: an efficient directional multiresolution image representation. *IEEE Trans. Image Processing*, to appear.

[8] D. L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.

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

[10] D. L. Donoho and I. M. Johnston. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[11] M. Elad. Why simple shrinkage is still relevant for redundant representations? *IEEE Trans. On Information Theory*, submitted.

[12] G. Fan and X. Xia. Image denoising using local contextual hidden markov model in the wavelet domain. *IEEE Signal Processing Letters*, 8(5):125–128, May 2001.

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

[14] J. Hurri, A. Hyv, R. Karhunen, and E. Oja. Wavelets and natural image statistics. In *Proc. Scandinavian Conf. on Image Analysis '97*, Lappenranta, Finland, 1998.

[15] X. Li and M. T. Orchard. Spatially adaptive image denoising under overcomplete expansion. In *Proc. IEEE ICIP*, pages 300–303, Vancouver, Canada, March 2000.

[16] S. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, July 1989.

[17] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[18] B. Matalon, M. Elad, and M. Zibulevsky. Image denoising with the contourlet transform. In *Proceedings of SPARSE'05*, Rennes, France, November 2005.

[19] P. Moulin and J. Liu. Analysis of multiresolution image denoising schemes using generalized-gaussian priors, October 1998.

[20] N. Nezamoddini-Kachouie, P. Fieguth, and E. Jernigan. Bayesshrink ridgelets for image denoising. In *Proc. ICIAR 2004*, Porto, Portugal, September 2004.

[21] A. Nosratinia. Denoising of jpeg images by re-application of jpeg. *Journal of VLSI Signal Processing*, 27(1):69–79, 2001.

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

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

[24] 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 Processing*, 11(5):545–557, May 2002.

[25] J. Portilla, V. Strela, M. Wainwright, and E. P. Simoncelli. Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Transactions on Image Processing*, 12(11):1338–1351, November 2003.

[26] D.L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Physics Review Letters*, 73(6):814–817, 1994.

[27] L. Sendur and I. W. Selesnick. Bivariate shrinkage with local variance estimation. *IEEE Signal Processing Letters*, 9(12):438–441, Dec. 2002.

[28] Z. Shan and S. Aviyente. Image denoising based on the wavelet co-occurrence matrix. In *Proc. IEEE ICASSP 2005*, pages 645–648, Philadelphia, USA, March 2005.

[29] E. P. Simoncelli. Bayesian denoising of visual images in the wavelet domain. In P Mller and B Vidakovic, editors, *Bayesian Inference in Wavelet Based Models*. Springer-Verlag, Lecture Notes in Statistics, 1999.

[30] E. P. Simoncelli. Modeling the joint statistics of images in the wavelet domain. In *Proc. SPIE, 44th Annual Meeting*, pages 188–195, Denver, CO, 1999.

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

[32] J.L. Starck, E.J. Candes, and D.L. Donoho. The curvelet transform for image denoising. *IEEE Trans. Image Processing*, 11(6):670–684, June 2002.

[33] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proc. IEEE international conference on computer vision*, Bombay, India, 1998.

[34] V. N. Vapnik. *Statistical Learning Theory*. John Wiley, 1998.