Image Denoising Via Sparse and Redundant Representations Over Learned Dictionaries

Michael Elad and Michal Aharon

***Abstract*—We address the image denoising problem, where zero-mean white and homogeneous Gaussian additive noise is to be removed from a given image. The approach taken is based on sparse and redundant representations over trained dictionaries. Using the K-SVD algorithm, we obtain a dictionary that describes the image content effectively. Two training options are considered: using the corrupted image itself, or training on a corpus of high-quality image database. Since the K-SVD is limited in handling small image patches, we extend its deployment to arbitrary image sizes by defining a global image prior that forces sparsity over patches in every location in the image. We show how such Bayesian treatment leads to a simple and effective denoising algorithm. This leads to a state-of-the-art denoising performance, equivalent and sometimes surpassing recently published leading alternative denoising methods.**

***Index Terms*—Bayesian reconstruction, dictionary learning, discrete cosine transform (DCT), image denoising, K-SVD, matching pursuit, maximum *a posteriori* (MAP) estimation, redundancy, sparse representations.**

I. Introduction

In THIS paper, we address the classic image denoising problem: An ideal image $\mathbf{x}$ is measured in the presence of an additive zero-mean white and homogeneous Gaussian noise, $\mathbf{v}$, with standard deviation σ. The measured image $\mathbf{y}$ is, thus

$$\mathbf{y} = \mathbf{x} + \mathbf{v}. \tag{1}$$

We desire to design an algorithm that can remove the noise from $\mathbf{y}$, getting as close as possible to the original image, $\mathbf{x}$.

The image denoising problem is important, not only because of the evident applications it serves. Being the simplest possible inverse problem, it provides a convenient platform over which image processing ideas and techniques can be assessed. Indeed, numerous contributions in the past 50 years or so addressed this problem from many and diverse points of view. Statistical estimators of all sorts, spatial adaptive filters, stochastic analysis, partial differential equations, transform-domain methods, splines and other approximation theory methods, morphological analysis, order statistics, and more, are some of the many directions explored in studying this problem. In this paper, we have no intention to provide a survey of this vast activity. Instead, we intend to concentrate on one specific approach towards the image denoising problem that we find to be highly effective and promising: the use of *sparse and redundant representations over trained dictionaries*.

Using redundant representations and sparsity as driving forces for denoising of signals has drawn a lot of research attention in the past decade or so. At first, sparsity of the unitary wavelet coefficients was considered, leading to the celebrated shrinkage algorithm [1]–[9]. One reason to turn to redundant representations was the desire to have the shift invariance property [10]. Also, with the growing realization that regular separable 1-D wavelets are inappropriate for handling images, several new tailored multiscale and directional redundant transforms were introduced, including the curvelet [11], [12], contourlet [13], [14], wedgelet [15], bandlet [16], [17], and the steerable wavelet [18], [19]. In parallel, the introduction of the matching pursuit [20], [21] and the basis pursuit denoising [22] gave rise to the ability to address the image denoising problem as a direct sparse decomposition technique over redundant dictionaries. All these lead to what is considered today as some of the best available image denoising methods (see [23]–[26] for few representative works).

While the work reported here is also built on the very same sparsity and redundancy concepts, it is adopting a different point of view, drawing from yet another recent line of work that studies example-based restoration. In addressing general inverse problems in image processing using the Bayesian approach, an image prior is necessary. Traditionally, this has been handled by choosing a prior based on some simplifying assumptions, such as spatial smoothness, low/max-entropy, or sparsity in some transform domain. While these common approaches lean on a guess of a mathematical expression for the image prior, the example-based techniques suggest to learn the prior from images somehow. For example, assuming a spatial smoothness-based Markov random field prior of a specific structure, one can still question (and, thus, train) the derivative filters to apply on the image, and the robust function to use in weighting these filters' outcome [27]–[29].

When this prior-learning idea is merged with sparsity and redundancy, it is the dictionary to be used that we target as the learned set of parameters. Instead of the deployment of a prechosen set of basis functions as the curvelet or contourlet would do, we propose to learn the dictionary from examples. In this work we consider two training options: 1) training the dictionary using patches from the corrupted image itself or 2) training on a corpus of patches taken from a high-quality set of images.

This idea of learning a dictionary that yields sparse representations for a set of training image-patches has been studied in a sequence of works [30]–[37]. In this paper, we propose the

Manuscript received December 18, 2005; revised May 3, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tamas Sziranyi.

The authors are with the Department of Computer Science, The Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: elad@cs.technion.ac.il; michalo@cs.technion.ac.il).

Digital Object Identifier 10.1109/TIP.2006.881969

K-SVD algorithm [36], [37] because of its simplicity and efficiency for this task. Also, due to its structure, we shall see how the training and the denoising fuse together naturally into one coherent and iterated process, when training is done on the given image directly.

Since dictionary learning is limited in handling small image patches, a natural difficulty arises: How can we use it for general images of arbitrary size? In this work, we propose a global image prior that forces sparsity over patches in every location in the image (with overlaps). This aligns with a similar idea, appearing in [29], for turning a local MRF-based prior into a global one. We define a maximum *a posteriori* probability (MAP) estimator as the minimizer of a well-defined global penalty term. Its numerical solution leads to a simple iterated patch-by-patch sparse coding and averaging algorithm that is closely related to the ideas explored in [38]–[40] and generalizes them.

When considering the available global and multiscale alternative denoising schemes (e.g., based on curvelet, contourlet, and steerable wavelet), it looks like there is much to be lost in working on small patches. Is there any chance of getting a comparable denoising performance with a local-sparsity based method? In that respect, the image denoising work reported in [23] is of great importance. Beyond the specific novel and highly effective algorithm described in that paper, Portilla and his coauthors posed a clear set of comparative experiments that standardize how image denoising algorithms should be assessed and compared one versus the other. We make use of these exact experiments and show that the newly proposed algorithm performs similarly, and, often, better, compared to the denoising performance reported in their work.

To summarize, the novelty of this paper includes the way we use local sparsity and redundancy as ingredients in a global Bayesian objective—this part is described in Section II, along with its emerging iterated numerical solver. Also novel in this work is the idea to train dictionaries for the denoising task, rather than use prechosen ones. As already mentioned earlier, when training is done on the corrupted image directly, the overall training-denoising algorithm becomes fused into one iterative procedure that comprises of steps of denoising of the image, followed by an update of the dictionary. This is described in Section III in detail. In Section IV, we show some experimental results that demonstrate the effectiveness of this algorithm.

II. From Local to Global Bayesian Reconstruction

In this section, we start the presentation of the proposed denoising algorithm by first introducing how sparsity and redundancy are brought to use. We do that via the introduction of the *Sparseland* model. Once this is set, we will discuss how local treatment on image patches turns into a global prior in a Bayesian reconstruction framework.

A. Sparseland Model for Image Patches

We consider image patches of size $\sqrt{n} \times \sqrt{n}$ pixels, ordered lexicographically as column vectors $\mathbf{x} \in \mathcal{R}^n$. For the construction of the *Sparseland* model, we need to define a dictionary (matrix) of size $\mathbf{D} \in \mathcal{R}^{n \times k}$ (with $k > n$, implying that it is redundant). At the moment, we shall assume that this matrix is known and fixed. Put loosely, the proposed model suggests that every image patch, $\mathbf{x}$, could be represented sparsely over this dictionary, i.e., the solution of

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} ||\boldsymbol{\alpha}||_0 \text{ subject to } \mathbf{D}\boldsymbol{\alpha} \approx \mathbf{x} \tag{2}$$

is indeed very sparse, $||\hat{\boldsymbol{\alpha}}||_0 \ll n$. The notation $||\boldsymbol{\alpha}||_0$ stands for the count of the nonzero entries in $\boldsymbol{\alpha}$. The basic idea here is that every signal instance from the family we consider can be represented as a linear combination of few columns (atoms) from the redundant dictionary $\mathbf{D}$.

This model should be made more precise by replacing the rough constraint $\mathbf{D}\boldsymbol{\alpha} \approx \mathbf{x}$ with a clear requirement to allow a bounded representation error, $||\mathbf{D}\boldsymbol{\alpha} - \mathbf{x}||_2 \leq \epsilon$. Also, one needs to define how deep is the required sparsity, adding a requirement of the form $||\hat{\boldsymbol{\alpha}}||_0 \leq L \ll n$, that states that the sparse representation uses no more than L atoms from the dictionary for every image patch instance. Alternatively, a probabilistic characterization can be given, defining the probability to obtain a representation with $||\hat{\boldsymbol{\alpha}}||_0$ nonzeros as a decaying function of some sort. Considering the simpler option between the two, with the triplet $(\epsilon, L, \mathbf{D})$ in place, our model is well defined.

Now assume that $\mathbf{x}$ indeed belongs to the $(\epsilon, L, \mathbf{D})$-*Sparseland* signals. Consider a noisy version of it, $\mathbf{y}$, contaminated by an additive zero-mean white Gaussian noise with standard deviation σ. The MAP estimator for denoising this image patch is built by solving

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} ||\boldsymbol{\alpha}||_0 \text{ subject to } ||\mathbf{D}\boldsymbol{\alpha} - \mathbf{y}||_2^2 \leq T \tag{3}$$

where T is dictated by ϵ, σ. The denoised image is, thus, given by $\hat{\mathbf{x}} = \mathbf{D}\hat{\boldsymbol{\alpha}}$ [22], [41], [42]. Notice that the above optimization task can be changed to be

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} ||\mathbf{D}\boldsymbol{\alpha} - \mathbf{y}||_2^2 + \mu||\boldsymbol{\alpha}||_0 \tag{4}$$

so that the constraint becomes a penalty. For a proper choice of μ, the two problems are equivalent. We will use this alternative terminology from now on, as it makes the presentation of later parts simpler to follow.

While this problem is, in general, very hard to solve, the matching and the basis pursuit algorithms can be used quite effectively [20]–[22] to get an approximated solution. Recent work established that those approximation techniques can be quite accurate if the solution is sparse enough to begin with [41], [42]. In this work, we will make use mainly of the orthonormal matching pursuit (OMP) because of its simplicity [21] and efficiency.

B. From Local Analysis to a Global Prior

If we want to handle a larger image $\mathbf{X}$ of size $\sqrt{N} \times \sqrt{N} (N \gg n)$, and we are still interested in using the above described model, one option is to redefine the model with a larger dictionary. Indeed, when using this model with a dictionary emerging from the contourlet or curvelet transforms, such scaling is simple and natural [26].

However, when we insist on using a specific fixed and small size dictionary $\mathbf{D} \in \mathcal{R}^{n \times k}$, this option no longer exists. Thus,

a natural question arises concerning the use of such a small dictionary in the first place. Two reasons come to mind: 1) when training takes place (as we will show in the next section), only small dictionaries can be composed; and furthermore; 2) a small dictionary implies a locality of the resulting algorithms, which simplifies the overall image treatment.

We next describe possible ways to use such a small dictionary when treating a large image. A heuristic approach is to work on smaller patches of size $\sqrt{n} \times \sqrt{n}$ and tile the results. In doing so, visible artifacts may occur on block boundaries. One could also propose to work on overlapping patches and average the results in order to prevent such blockiness artifacts, as, indeed, practiced in [38]–[40]. As we shall see next, a systematic global approach towards this problem leads to this very option as a core ingredient in an overall algorithm.

If our knowledge on the unknown large image $\mathbf{X}$ is fully expressed in the fact that every patch in it belongs to the $(\epsilon, L, \mathbf{D})$-*Sparseland* model, then the natural generalization of the above MAP estimator is the replacement of (4) with

$$\{\hat{\boldsymbol{\alpha}}_{ij}, \hat{\mathbf{X}}\} = \arg\min_{\boldsymbol{\alpha}_{ij}, \mathbf{X}} \lambda ||\mathbf{X} - \mathbf{Y}||_2^2 + \sum_{ij} \mu_{ij} ||\boldsymbol{\alpha}_{ij}||_0 + \sum_{ij} ||\mathbf{D}\boldsymbol{\alpha}_{ij} - \mathbf{R}_{ij}\mathbf{X}||_2^2. \quad (5)$$

In this expression, the first term is the log-likelihood global force that demands the proximity between the measured image, $\mathbf{Y}$, and its denoised (and unknown) version $\mathbf{X}$. Put as a constraint, this penalty would have read $||\mathbf{X} - \mathbf{Y}||_2^2 \leq \text{Const} \cdot \sigma^2$, and this reflects the direct relationship between λ and σ.

The second and the third terms are the image prior that makes sure that in the constructed image, $\mathbf{X}$, every patch $\mathbf{x}_{ij} = \mathbf{R}_{ij}\mathbf{X}$ of size $\sqrt{n} \times \sqrt{n}$ in every location (thus, the summation by i, j) has a sparse representation with bounded error. Similar conversion has also been practiced by Roth and Black when handling an MRF prior [29].

The matrix $\mathbf{R}_{ij}$ is an $n \times N$ matrix that extracts the (ij) block from the image. For an $\sqrt{N} \times \sqrt{N}$ image $\mathbf{X}$, the summation over i, j includes $(\sqrt{N} - \sqrt{n} + 1)^2$ items, considering all image patches of size $\sqrt{n} \times \sqrt{n}$ in $\mathbf{X}$ with overlaps. As to the coefficients μ_{ij}, those must be location dependent, so as to comply with a set of constraints of the form $||\mathbf{D}\boldsymbol{\alpha}_{ij} - \mathbf{x}_{ij}||_2^2 \leq T$.

C. Numerical Solution

When the underlying dictionary $\mathbf{D}$ is assumed known, the proposed penalty term in (5) has two kinds of unknowns: the sparse representations $\hat{\boldsymbol{\alpha}}_{ij}$ per each location, and the overall output image $\mathbf{X}$. Instead of addressing both together, we propose a block-coordinate minimization algorithm that starts with an initialization $\mathbf{X} = \mathbf{Y}$, and then seeks the optimal $\hat{\boldsymbol{\alpha}}_{ij}$. In doing so, we get a complete decoupling of the minimization task to many smaller ones, each of the form

$$\hat{\boldsymbol{\alpha}}_{ij} = \arg\min_{\boldsymbol{\alpha}} \mu_{ij} ||\boldsymbol{\alpha}||_0 + ||\mathbf{D}\boldsymbol{\alpha} - \mathbf{x}_{ij}||_2^2 \quad (6)$$

handling one image patch. Solving this using the orthonormal matching pursuit [21] is easy, gathering one atom at a time, and stopping when the error $||\mathbf{D}\boldsymbol{\alpha} - \mathbf{x}_{ij}||_2^2$ goes below T. This way, the choice of μ_{ij} has been handled implicitly. Thus, this stage works as a sliding window sparse coding stage, operated on each block of $\sqrt{n} \times \sqrt{n}$ at a time.

Given all $\hat{\boldsymbol{\alpha}}_{ij}$, we can now fix those and turn to update $\mathbf{X}$. Returning to (5), we need to solve

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{x}} \lambda ||\mathbf{X} - \mathbf{Y}||_2^2 + \sum_{ij} ||\mathbf{D}\hat{\boldsymbol{\alpha}}_{ij} - \mathbf{R}_{ij}\mathbf{X}||_2^2. \quad (7)$$

This is a simple quadratic term that has a closed-form solution of the form

$$\hat{\mathbf{X}} = \left(\lambda \mathbf{I} + \sum_{ij} \mathbf{R}_{ij}^T \mathbf{R}_{ij}\right)^{-1} \left(\lambda \mathbf{Y} + \sum_{ij} \mathbf{R}_{ij}^T \mathbf{D}\hat{\boldsymbol{\alpha}}_{ij}\right). \quad (8)$$

This rather cumbersome expression may mislead, as all it says is that averaging of the denoised patches is to be done, with some relaxation obtained by averaging with the original noisy image. The matrix to invert in the above expression is a diagonal one, and, thus, the calculation of (8) can be also done on a pixel-by-pixel basis, following the previously described sliding window sparse coding steps.

So far, we have seen that the obtained denoising algorithm calls for sparse coding of small patches, and an averaging of their outcomes. However, if minimization of (5) is our goal, then this process should proceed. Given the updated $\mathbf{X}$, we can repeat the sparse coding stage, this time working on patches from the already denoised image. Once this is done, a new averaging should be calculated, and so on, and so forth. Thus, we obtain exactly what Guleryuz suggested in his work—iterated denoising via sparse representation, and we may regard the analysis proposed here as a rigorous way to justify such iterated scheme [38]–[40].

III. Example-Based Sparsity and Redundancy

The entire discussion so far has been based on the assumption that the dictionary $\mathbf{D} \in \mathcal{R}^{n \times k}$ is known. We can certainly make some educated guesses as to which dictionaries to use. In fact, following Guleryuz's work, the DCT seems like such a plausible choice [38]–[40]. Indeed, we might do better by using a redundant version of the DCT,[1] as practiced in [36]. Still, the question remains: Can we make a better choice for $\mathbf{D}$ based on training? We now turn to discuss this option. We start with the simpler (and less effective) option of training the dictionary on a set of image patches taken from good quality images, and then turn to discuss the option of training on the corrupted image itself.

A. Training on the Corpus of Image Patches

Given a set of image patches $\mathcal{Z} = \{\mathbf{z}_j\}_{j=1}^M$, each of size $\sqrt{n} \times \sqrt{n}$, and assuming that they emerge from a specific $(\epsilon, L, \mathbf{D})$-*Sparseland* model, we would like to estimate this model parameters, $(\epsilon, L, \mathbf{D})$. Put formally, we seek the dictionary $\mathbf{D}$ that minimizes

$$\varepsilon\left(\mathbf{D}, \{\boldsymbol{\alpha}_j\}_{j=1}^M\right) = \sum_{j=1}^M \left[\mu_j ||\boldsymbol{\alpha}_j||_0 + ||\mathbf{D}\boldsymbol{\alpha}_j - \mathbf{z}_j||_2^2\right]. \quad (9)$$

Just as before, the above expression seeks to get a sparse representation per each of the examples in $\mathcal{Z}$, and obtain a small

[1]Such a version is created by using a redundant Fourier dictionary and a mirror extension of the signal to restrict the transform to real entries.

Task: Denoise a given image Y form white and additive Gaussian white noise with standard deviation σ.

Algorithm Parameters: n - block size, k - dictionary size, J - number of training iterations, λ - Lagrange multiplier, and C - noise gain.

$$\min_{\mathbf{X},\mathbf{D},\mathbf{A}} \left\{ \lambda\|\mathbf{Y}-\mathbf{X}\| + \sum_{ij} \mu_{ij}\|\alpha_{ij}\|_0 + \sum_{ij} \|\mathbf{D}\alpha_{ij} - R_{ij}X\|_2^2 \right\}$$

1. Initialization : Set $\mathbf{X}=\mathbf{Y}$, $\mathbf{D}=$ overcomplete DCT dictionary.
2. Repeat J times:
   - *Sparse Coding Stage*: Use any pursuit algorithm to compute the representation vectors α_{ij} for each patch $R_{ij}\mathbf{X}$, by approximating the solution of
   $$\forall_{ij} \ \min_{\alpha_{ij}} \|\alpha_{ij}\|_0 \quad \text{s.t.} \quad \|Rij\mathbf{X} - \mathbf{D}\alpha_{ij}\|_2^2 \le (C\sigma)^2.$$
   - *Dictionary Update Stage*: For each column $l=1,2,\ldots,k$ in $\mathbf{D}$, update it by
     - Find the set of patches that use this atom, $\omega_l = \{(i,j)|\alpha_{ij}(l)\neq 0\}$.
     - For each index $(i,j)\in\omega_l$, compute its representation error
     $$\mathbf{e}_{ij}^l = R_{ij}\mathbf{X}_{ij} - \sum_{m\neq l} \mathbf{d}_m \alpha_{ij}(m).$$
     - set $\mathbf{E}_l$ as the matrix whose columns are $\{\mathbf{e}_{ij}^l\}_{(i,j)\in\omega_l}$
     - Apply SVD decomposition $\mathbf{E}_l = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T$. Choose the updated dictionary column $\tilde{\mathbf{d}}_l$ to be the first column of $\mathbf{U}$. Update the coefficient values $\{\alpha_{ij}(l)\}_{(i,j)\in\omega_l}$ to be the entries of $\mathbf{V}$ multiplied by $\boldsymbol{\Delta}(1,1)$.
3. Set:
$$X = \left(\lambda\mathbf{I} + \sum_{ij} R_{ij}^T R_{ij}\right)^{-1} \left(\lambda\mathbf{Y} + \sum_{ij} R_{ij}^T \mathbf{D}\alpha_{ij}\right)$$

Fig. 1. Denoising procedure using a dictionary trained on patches from the corrupted image. For our experiments, we used the OMP pursuit method, and set $J=10, \lambda = 30/\sigma$ and $C=1.15$.

representation error. The choice for μ_j dictates how those two forces should be weighted, so as to make one of them a clear constraint. For example, constraining $\forall j \|\boldsymbol{\alpha}_j\|_0 = L$ implies specific values for μ_j, while requiring $\forall j \|\mathbf{D}\boldsymbol{\alpha}_j - \mathbf{z}_j\|_2^2 \le \epsilon^2$ leads to others.

The K-SVD proposes an iterative algorithm designed to handle the above task effectively [36], [37]. Adopting again the block-coordinate descent idea, the computations of $\mathbf{D}$ and $\{\boldsymbol{\alpha}_j\}_{j=1}^M$ are separated. Assuming that $\mathbf{D}$ is known, the penalty posed in (9) reduces to a set of M sparse coding operations, very much like the ones seen in (6). Thus, OMP can be used again to obtain the near-optimal (recall that OMP is an approximation algorithm, and, thus, a true minimization is not guaranteed) set of representation vectors $\{\boldsymbol{\alpha}_j\}_{j=1}^M$.

Assuming these representation vectors fixed, the K-SVD proposes an update of the dictionary one column at a time. As it turns out, this update can be done optimally, leading to the need to perform a SVD operation on residual data matrices, computed only on the examples that use this atom. This way, the value of $\varepsilon(\mathbf{D}, \{\boldsymbol{\alpha}_j\}_{j=1}^M)$ is guaranteed to drop per an update of each dictionary atom, and along with this update, the representation coefficients change as well (see [36] and [37] for more details).

When adopted to the denoising task at hand, a crucial step is the choice of the examples to train on. Is there really a universal dictionary that fits all images well? If there is one, which examples shall we use to find it? The experiments that follow in the next section bring us to the conclusion that while a reasonably good dictionary that fits all is indeed within reach, extracting state-of-the-art denoising performance calls for a more complex model that uses several dictionaries switched by content—an option we do not explore in this work.

Also, since the penalty minimized here in (9) is a highly nonconvex functional, local minimum solutions are likely to haunt us. Thus, a wise initialization could be of great worth. In our

TABLE I
SUMMARY OF THE DENOISING PSNR RESULTS IN DECIBELS. IN EACH CELL. FOUR DENOISING RESULTS ARE REPORTED. TOP LEFT: RESULTS OF PORTILLA *ET AL.* [23]. TOP RIGHT: OVERCOMPLETE DCT. BOTTOM LEFT: GLOBAL TRAINED DICTIONARY. BOTTOM RIGHT: ADAPTIVE DICTIONARY TRAINED ON NOISY IMAGE. IN EACH SUCH SET WE HIGHLIGHTED THE BEST RESULT. THE THREE LATTER METHODS WERE EXECUTED WITH $\lambda = 30/\sigma$. ALL NUMBERS ARE AN AVERAGE OVER FIVE EXPERIMENTS. THE LAST TWO COLUMNS PRESENT THE AVERAGE RESULTS OVER ALL IMAGES AND THEIR VARIANCE

$\sigma/PSNR$	Lena		Barb		Boats		Fgrpt		House		Peppers		Average		σ_{PSNR}	
2/42.11	43.23	43.55	43.29	43.61	42.99	43.07	**43.05**	42.92	44.07	44.38	43.00	43.30	43.27	43.47	0.012	0.017
	43.23	**43.58**	43.10	**43.67**	41.86	**43.14**	42.94	42.99	44.27	**44.47**	42.90	**43.33**	43.05	**43.53**	0.018	0.017
5/34.15	38.49	38.51	37.79	37.93	36.97	37.09	**36.68**	36.48	38.65	39.07	37.31	37.67	37.65	37.79	0.014	0.016
	38.48	**38.60**	37.32	**38.08**	36.64	**37.22**	36.56	36.65	38.86	**39.37**	37.65	**37.78**	37.59	**37.95**	0.016	0.017
10/28.13	**35.61**	35.28	34.03	33.97	33.58	33.44	**32.45**	32.14	35.35	35.41	33.77	33.93	34.13	34.03	0.017	0.026
	35.40	35.47	33.07	**34.42**	33.53	**33.64**	32.23	32.39	35.69	**35.98**	**34.32**	34.28	34.04	**34.36**	0.024	0.027
15/24.61	**33.90**	33.38	31.86	31.63	31.70	31.38	**30.14**	29.71	33.64	33.49	31.74	31.76	32.16	31.89	0.024	0.032
	33.60	33.70	30.61	**32.37**	31.63	**31.73**	29.86	30.06	34.03	**34.32**	**32.37**	32.22	32.02	**32.40**	0.030	0.035
20/22.11	**32.66**	32.00	30.32	29.95	**30.38**	29.91	**28.60**	28.01	32.39	32.17	30.31	30.20	30.78	30.37	0.031	0.024
	32.27	32.38	28.87	**30.83**	30.24	30.36	28.21	28.47	32.88	**33.20**	**30.92**	30.82	30.57	**31.01**	0.025	0.027
25/20.17	**31.69**	30.89	29.13	28.65	**29.37**	28.78	**27.45**	26.65	31.40	31.03	29.21	29.01	29.71	29.17	0.037	0.037
	31.20	31.32	27.57	**29.60**	29.17	29.28	26.94	27.26	31.82	**32.15**	**29.84**	29.73	29.42	**29.89**	0.035	0.036
50/14.15	**28.61**	27.44	**25.48**	24.75	**26.38**	25.57	**24.16**	22.01	**28.26**	27.41	25.90	25.25	**26.47**	25.41	0.049	0.049
	27.77	27.79	24.06	25.47	25.91	25.95	22.68	23.24	27.91	27.95	26.12	**26.13**	25.74	26.01	0.051	0.058
75/10.63	**26.84**	25.63	**23.65**	22.83	**24.79**	23.85	**22.40**	19.28	**26.41**	25.10	**24.00**	23.12	**24.68**	23.3	0.061	0.053
	25.81	25.80	22.54	23.01	24.02	23.98	19.73	19.97	25.33	25.22	23.78	23.69	23.54	23.61	0.070	0.060
100/8.13	**25.64**	24.42	**22.61**	21.89	**23.75**	22.79	**21.22**	17.99	**25.11**	23.78	**22.66**	21.55	**23.50**	22.07	0.070	0.044
	24.45	24.46	21.73	21.89	22.83	22.81	18.23	18.30	23.86	23.71	21.88	21.75	22.16	22.15	0.050	0.046

experiments we started with the already mentioned redundant DCT, which proves to be a good dictionary choice. This also enabled us to apply fewer number of iterations.

Another puzzling issue is the redundancy factor k/n—How should we choose k, the number of columns in $\mathbf{D}$? Is there an optimal choice? In this work, we do not address this important question, and simply choose a value we find empirically to perform well. Further work is required to explore this matter.

B. Training on the Corrupted Image

Instead of supplying an artificial set of examples to train on, as proposed above, one could take the patches from the corrupted image, $\mathcal{Z} = \{\mathbf{y}_j\}_{j=1}^{M}$, where $M = (\sqrt{N} - \sqrt{n} + 1)^2$. Since the K-SVD dictionary learning process has in it a noise rejection capability (see experiments reported in [36]), this seems like a natural idea. Furthermore, rather than using unrelated examples that call for the universality assumption of the *Sparseland* model, this option tailors the dictionary to the image to be treated.

At first sight, this change in the origin of the examples to train on seems to be of technical worth, and has no impact on the overall algorithm. However, a close inspection of both the functional $\varepsilon(\mathbf{D}, \{\boldsymbol{\alpha}_j\}_{j=1}^{M})$ in (9), and the global MAP penalty in (5), reveals the close resemblance between the two. This implies that the dictionary design could be embedded within the Bayesian approach. Returning to (5), we can regard also $\mathbf{D}$ as an unknown, and define our problem as

$$\{\hat{\mathbf{D}}, \hat{\boldsymbol{\alpha}}_{ij}, \hat{\mathbf{X}}\} = \arg\min_{\mathbf{D}, \boldsymbol{\alpha}_{ij}, \mathbf{X}} \lambda \|\mathbf{X} - \mathbf{Y}\|_2^2 + \sum_{ij} \mu_{ij} \|\boldsymbol{\alpha}_{ij}\|_0 + \sum_{ij} \|\mathbf{D}\boldsymbol{\alpha}_{ij} - \mathbf{R}_{ij}\mathbf{X}\|_2^2. \quad (10)$$

Following the previously constructed algorithm, we can assume a fixed $\mathbf{D}$ and $\mathbf{X}$, and compute the representations $\hat{\boldsymbol{\alpha}}_{ij}$. This requires, as before, a sparse coding stage that deploys the OMP. Given those representations, the dictionary can be now updated, using a sequence of K-SVD operations.

Once done, the output image can be computed, using (8). However, an update of the output image $\mathbf{X}$ changes the noise level σ, which up until now has been considered as known, and was used in the preceding two stages. Therefore, we choose to perform several more iterations of representation computation and dictionary update, using the same value of σ, before finding the output image $\mathbf{X}$. This algorithm is described in detail in Fig. 1.

In evaluating the computational complexity of this algorithm, we consider all three stages—sparse coding (OMP process), dictionary update (these stages are iterated J times), and final averaging process. All stages can be done efficiently, requiring $O(nkLJ)$ operations per pixel, where n is the block dimension, k is the number of atoms in the dictionary, and L is the

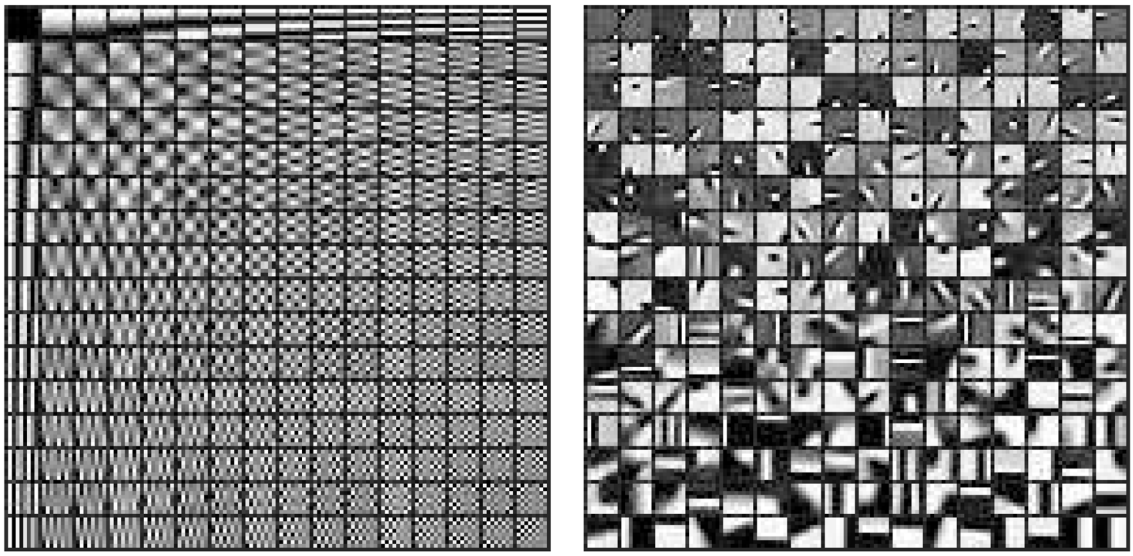

Fig. 2. Left: Overcomplete DCT dictionary. Right: Globally trained dictionary.

Fig. 3. Sample from the images used for training the global dictionary.

number of nonzero elements in each coefficient vector. L depends strongly on the noise level, e.g., for $\sigma = 10$, the average L is 2.96, and for $\sigma = 20$, the average L is 1.12.

IV. Results

In this section, we demonstrate the results achieved by applying the above methods on several test images, and with several dictionaries. The tested images, as also the tested noise levels, are all the same ones as those used in the denoising experiments reported in [23], in order to enable a fair comparison.

Table I summarizes these denoising results for the DCT dictionary, the globally trained dictionary, and training on the corrupted images directly (referred to hereafter as the adaptive dictionary). In all this set of experiments, the dictionaries used were of size 64×256, designed to handle image patches of size 8×8 pixels $(n = 64, k = 256)$. Every result reported is an average over 5 experiments, having different realizations of the noise.

The redundant DCT dictionary is described on the left side of Fig. 2, each of its atoms shown as an 8×8 pixel image. This dictionary was also used as the initialization for all the training algorithms that follow. The globally trained dictionary is shown on the right side of Fig. 2. This dictionary was produced by the K-SVD algorithm (executed 180 iterations, using OMP for sparse coding with $L = 6$), trained on a data-set of 100 000 8×8 patches. Those patches were taken from an arbitrary set of clean natural images (unrelated to the test images), some of which are shown in Fig. 3.

In all experiments, the denoising process included a sparse-coding of each patch of size 8×8 pixels from the noisy image. Using the OMP, atoms were accumulated till the average error passed the threshold, chosen empirically to be $\epsilon = 1.15 \cdot \sigma$. This means that our algorithm assumes the knowledge of σ—very much like that assumed in [23]. The denoised patches were averaged, as described in (8), using $\lambda = 30/\sigma$ (see below for an explanation for this choice of λ). We chose to apply only one iteration in the iterative process suggested in Section II-C. Following iterations requires knowledge of the new noisy parameter σ, which is unknown after first changing $\mathbf{X}$.

When training the dictionary on overlapping patches from the noisy image itself, each such experiment included $(256 - 7)^2 = 62\,001$ patches (all available patches from the 256×256 images, and every second patch from every second row in the 512×512 size images). The algorithm described in detail in Fig. 1 was applied.

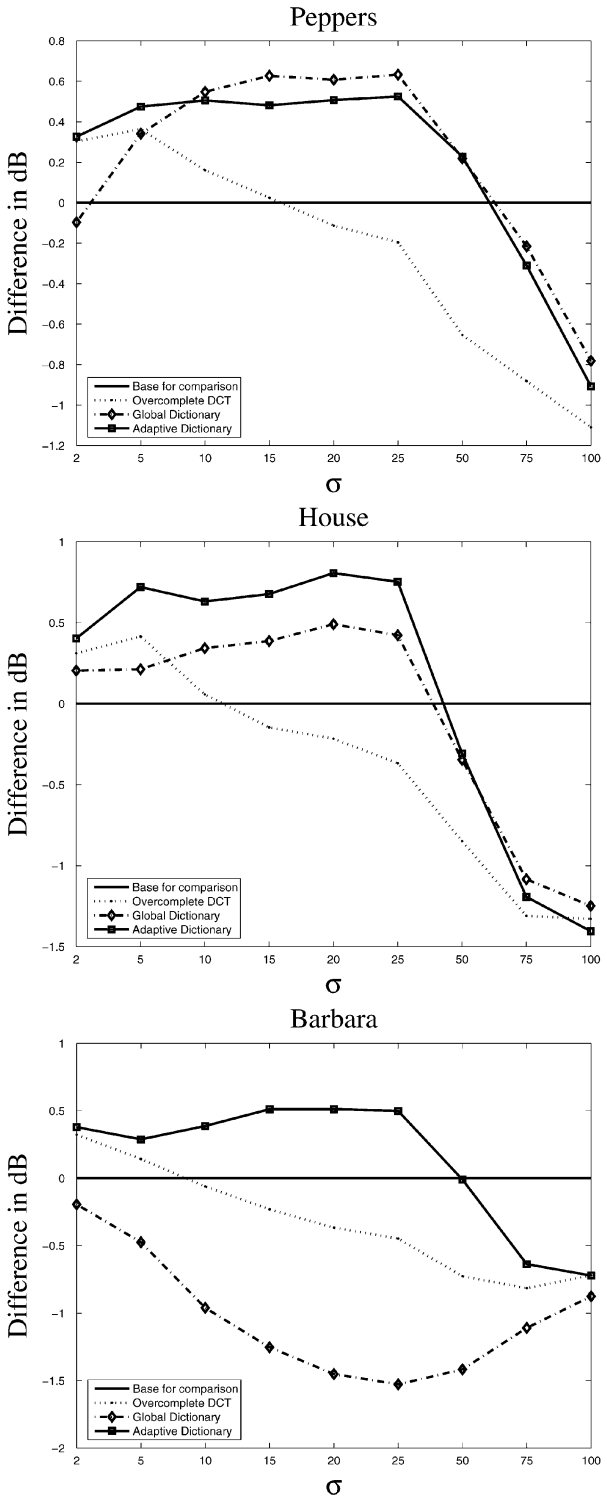

Fig. 4. Comparison between the three presented methods (overcomplete DCT, global trained dictionary, and adaptive dictionary trained on patches from the noisy image) and the results achieved recently in [23] for three test images.

As can be seen from Table I, the results of all methods are very close to each other in general. Averaging the results that correspond to [23] in this table for noise levels lower than,[2] $\sigma = 50$ the value is 34.62 dB. A similar averaging over the DCT dictionary results gives 34.45 dB, implying an average difference of 0.17 dB, in favor of Portilla's method. This is the same case with the globally trained dictionary, which means that our attempt to train one global dictionary for images performs as good as the fixed redundant DCT. However, for the method of the image-adaptive dictionary, an average of 34.86 dB is obtained, giving an average advantage of 0.24 dB over Portilla's method. For the higher noise power experiments, our approach deteriorates faster and achieves weaker results.

[2]The strong noise experiments are problematic to analyze, because clipping of the dynamic range to [0, 255], as often done, causes a severe deviation from the Gaussian distribution model assumed.

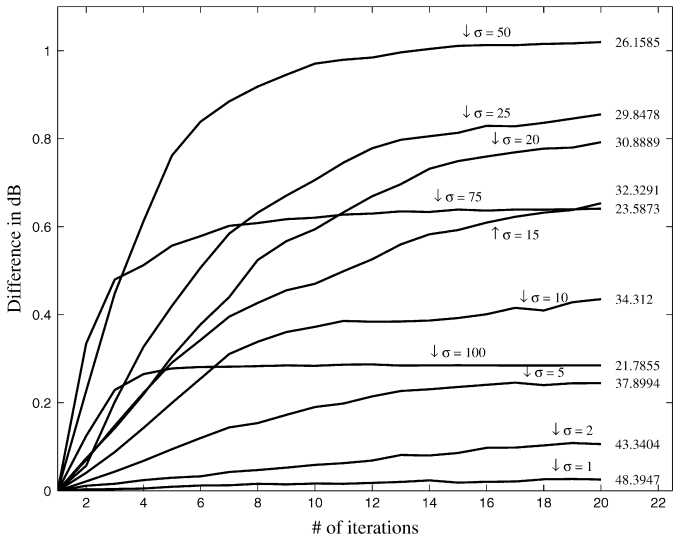

Fig. 5. Improvement in the denoising results after each iteration of the K-SVD algorithm, executed on noisy patches of the image "Peppers."

In order to better visualize the results and their comparison to those in [23], Fig. 4 presents the difference of the denoising results of the two proposed methods and the overcomplete DCT compared with those of [23] (which appears as a zero straight reference line). This comparison is presented for the images "Peppers," "House," and "Barbara." Notice that, for these images, the adaptive dictionary outperforms the reported results of Portilla *et al.* for all noise levels lower than $\sigma = 50$, while the global dictionary often achieves very close results. In the image "Barbara," however, which contains high-frequency texture areas, the adaptive dictionary that learns the specific characteristics has a clear advantage over the globally trained dictionary.

Fig. 5 further describes the behavior of the denoising algorithm that uses the adaptive dictionary. Each K-SVD iteration improves the denoising results, with the initial dictionary set to be the overcomplete DCT. A graph presenting this consistent improvement for several noise levels is presented in Fig. 5. All graphs show the improvement over the first iteration, and, therefore, all curves start at zero, going towards positive values. As can be seen, a gain of up to 1 dB is achievable. Fig. 6 shows the results of the proposed algorithms for the image "Barbara," and for $\sigma = 20$. The final adaptive dictionary that leads to those results is presented in Fig. 7.

We now turn to study the effect of the parameter λ in (8). As expected, we found that a proper choice for λ is dependent on the noise level. As the noise increases, better results are achieved

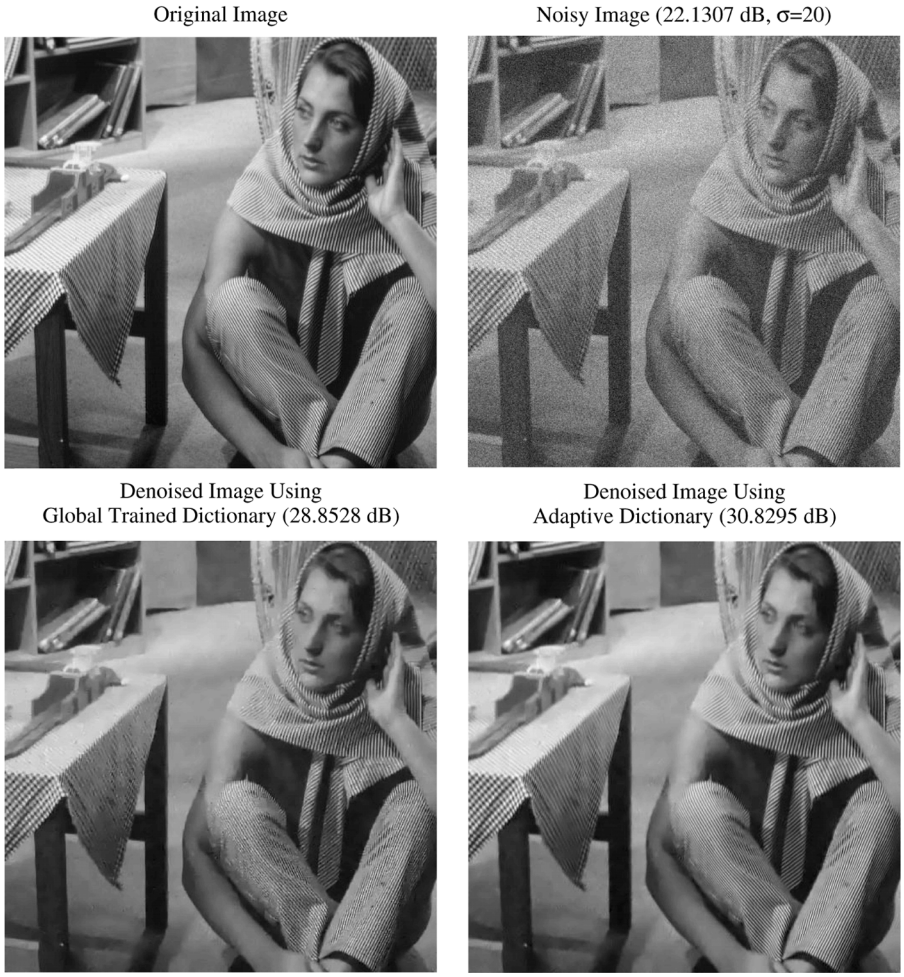

Fig. 6. Example of the denoising results for the image "Barbara" with $\sigma = 20$—the original, the noisy, and two restoration results.

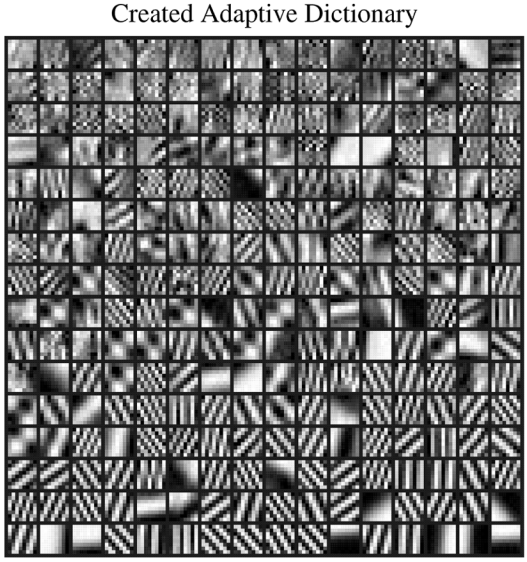

Fig. 7. Example of the denoising results for the image "Barbara" with $\sigma = 20$—the adaptively trained dictionary.

with small values of λ, and vice versa. This is, indeed, expected, as relatively "clean" images should have a stronger effect on the outcome, while very noisy ones should effect the outcome weakly, if at all. We tested several values for this parameter, and found empirically that the best results are achieved with $\lambda \approx 30/\sigma$. It is interesting to see that all three denoising methods (overcomplete DCT, global dictionary, and adaptive dictionary trained on noisy patches), and all noise levels generally agree with this choice. In Fig. 8, we present the improvement (and later, deterioration) achieved when increasing the value of λ in the averaging process (8). In Fig. 8, one image ("Peppers") was tested with four noise levels $(\sigma = 5, 10, 20, 50)$ and with all three methods, resulting with 12 curves. The choice $\lambda = 30/\sigma$ seems to be near the peak for all these graphs.

To conclude this experimental section, we refer to our arbitrary choice of $k = 256$ dictionary atoms (this choice had an effect over all three experimented methods). We conducted another experiment, which compares between several values of k. In this experiment, we tested the denoising results of the three proposed methods on the image "House" for an initial noise level of $\sigma = 15$ (24.61 dB) and $\lambda = 30/\sigma$. The tested redundancy values (of k) were 64, 128, 256, and 512. The average results of four executions (per each test) are presented in Fig. 9. As can be seen, the increase of the number of dictionary elements generally improves the results, although this improvement is small (0–0.16 dB). This increase is most effective in the adaptive dictionary method.

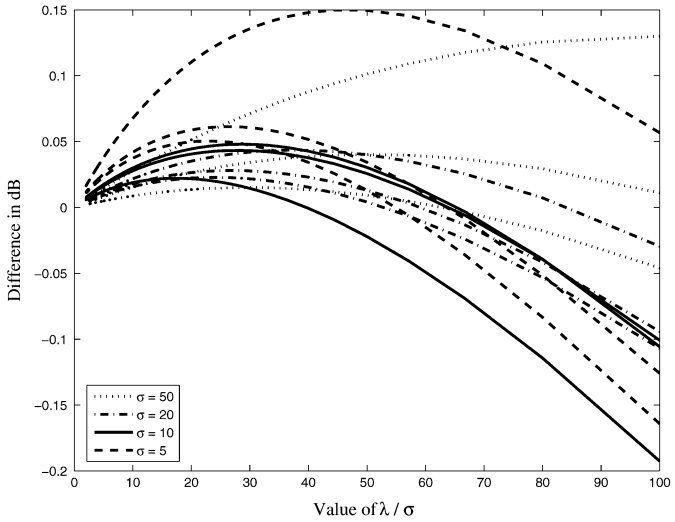

Fig. 8. Improvement (and later, deterioration) of the denoising results when increasing the value of λ in the averaging process in (8).

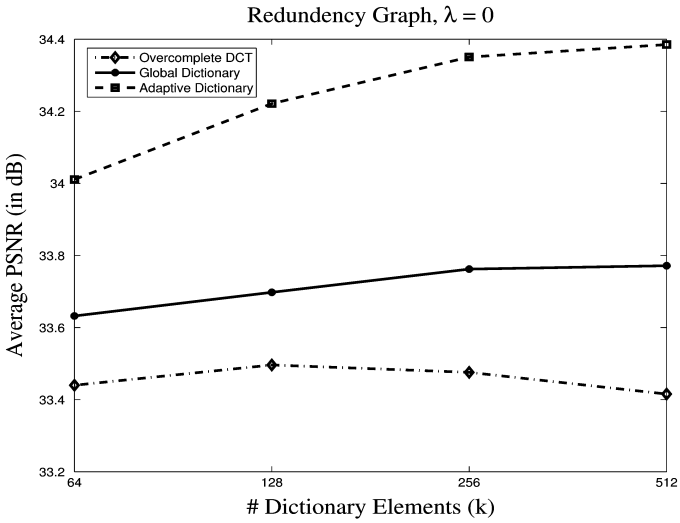

Fig. 9. Effect of changing the number of dictionary elements (k) on the final denoising results for the image "House" and for $\sigma = 15$.

V. Conclusion and Further Work

This work has presented a simple method for image denoising, leading to state-of-the-art performance, equivalent to and sometimes surpassing recently published leading alternatives. The proposed method is based on local operations and involves sparse decompositions of each image block under one fixed over-complete dictionary, and a simple average calculations. The content of the dictionary is of prime importance for the denoising process—we have shown that a dictionary trained for natural real images, as well as an adaptive dictionary trained on patches of the noisy image itself, both perform very well.

There are several research directions that we are currently considering, such as using several dictionaries and switching between them by content, optimizing the parameters, replacing the OMP by a better pursuit technique, and more. Beyond these, one direction we consider to be promising is a generalization to multiscale dictionaries. This work concentrated on small image patches, completely overlooking the global structure of the image, and the multiscale analysis that other techniques have exploited rather well. We are studying ways to extend this work to multiscale dictionaries, as it is clear that K-SVD cannot be directly deployed on larger blocks.

References

[1] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, Sep. 1994.
[2] D. L. Donoho, "De-noising by soft thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.
[3] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage—Asymptopia," *J. Roy. Statist. Soc. B—Methodological*, vol. 57, no. 2, pp. 301–337, 1995.
[4] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, no. 432, pp. 1200–1224, Dec. 1995.
[5] D. L. Donoho and I. M. Johnstone, "Minimax estimation via wavelet shrinkage," *Ann. Statist.*, vol. 26, no. 3, pp. 879–921, Jun. 1998.
[6] E. P. Simoncelli and E. H. Adelson, "Noise removal via Bayesian wavelet coring," in *Proc. Int. Conf. Image Processing*, Laussanne, Switzerland, Sep. 1996.
[7] A. Chambolle, R. A. DeVore, N.-Y. Lee, and B. J. Lucier, "Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. Image Process.*, vol. 7, no. 3, pp. 319–335, Mar. 1998.
[8] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 909–919, Apr. 1999.
[9] M. Jansen, *Noise Reduction by Wavelet Thresholding*. New York: Springer-Verlag, 2001.
[10] R. Coifman and D. L. Donoho, "Translation invariant de-noising," in *In Wavelets and Statistics, Lecture Notes in Statistics*. New York: Springer-Verlag, 1995, pp. 125–150, 1995.
[11] E. J. Candes and D. L. Donoho, "Recovering edges in ill-posed inverse problems: Optimality of curvelet frames," *Ann. Statist.*, vol. 30, no. 3, pp. 784–842, Jun. 2002.
[12] E. J. Candès and D. L. Donoho, "New tight frames of curvelets and the problem of approximating piecewise C^2 images with piecewise C^2 edges," *Commun. Pure Appl. Math.*, vol. 57, pp. 219–266, Feb. 2004.
[13] M. N. Do and M. Vetterli, *Contourlets, Beyond Wavelets*, G. V. Welland, Ed. New York: Academic, 2003.
[14] M. N. Do and M. Vetterli, "Framing pyramids," *IEEE Trans. Signal Process.*, vol. 51, pp. 2329–2342, Sep. 2003.
[15] D. L. Donoho, "Wedgelets: Nearly minimax estimation of edges," *Ann. Statist.*, vol. 27, no. 3, pp. 859–897, Jun. 1998.
[16] S. Mallat and E. LePennec, "Sparse geometric image representation with bandelets," *IEEE Trans. Image Process.*, vol. 14, no. 4, pp. 423–438, Apr. 2005.
[17] S. Mallat and E. LePennec, "Bandelet image approximation and compression," *SIAM J. Multiscale Model. Simul.*, 2005, to be published.
[18] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Pattern Anal. Mach. Intell.*, vol. 13, no. 9, pp. 891–906, Sep. 1991.
[19] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. H. Heeger, "Shiftable multi-scale transforms," *IEEE Trans Inf. Theory*, vol. 38, no. 2, pp. 587–607, Mar. 1992.
[20] S. Mallat and Z. Zhang, "Matching pursuit in a time-frequency dictionary," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
[21] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," presented at the 27th Annu. Asilomar Conf. Signals, Systems, and Computers, 1993.
[22] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–59, 2001.
[23] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet doma," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.
[24] 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.
[25] R. Eslami and H. Radha, "Translation-invariant contourlet transform and its application to image denoising," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3362–3374, Nov. 2006.

[26] B. Matalon, M. Elad, and M. Zibulevsky, "Improved denoising of images using modeling of the redundant contourlet transform," presented at the SPIE Conf. Wavelets, Jul. 2005.
[27] S. C. Zhu and D. Mumford, "Prior learning and Gibbs reaction-diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 11, pp. 1236–1250, Nov. 1997.
[28] E. Haber and L. Tenorio, "Learning regularization functionals," *Inv. Probl.*, vol. 19, pp. 611–626, 2003.
[29] S. Roth and M. J. Black, "Fields of experts: A framework for learning image priors," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2005, vol. 2, pp. 860–867.
[30] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vis. Res.*, vol. 37, pp. 311–325, 1997.
[31] K. Engan, S. O. Aase, and J. H. Hakon-Husoy, "Method of optimal directions for frame design," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1999, vol. 5, pp. 2443–2446.
[32] K. Kreutz-Delgado and B. D. Rao, "Focuss-based dictionary learning algorithms," presented at the Wavelet Applications in Signal and Image Processing VIII, 2000.
[33] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neur. Comput.*, vol. 15, no. 2, pp. 349–396, 2003.
[34] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neur. Comput.*, vol. 12, pp. 337–365, 2000.
[35] L. Lesage, R. Gribonval, F. Bimbot, and L. Benaroya, "Learning unions of orthonormal bases with thresholded singular value decomposition," presented at the IEEE Intl Conf. Acoustics, Speech, and Signal Processing, 2005.
[36] M. Aharon, M. Elad, and A. M. Bruckstein, "The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, to be published.
[37] M. Aharon, M. Elad, and A. M. Bruckstein, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *J. Linear Algebra Appl*.
[38] O. G. Guleryuz, "Weighted overcomplete denoising," presented at the Asilomar Conf. Signals and Systems, Pacific Grove, CA, Nov. 2003.
[39] O. G. Guleryuz, "Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising: Part I—Theory," *IEEE Trans. Image Process.*, vol. 15, no. 3, pp. 539–553, Mar. 2005.
[40] O. G. Guleryuz, "Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising: Part II—Adaptive algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 3, pp. 554–571, Mar. 2005.
[41] D. L. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.
[42] J. A. Tropp, "Just relax: Convex programming methods for subset selection and sparse approximation," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 1030–1051, Mar. 2005.

Michael Elad received the B.Sc., M.Sc., and D.Sc. degrees from the Department of Electrical Engineering, The Technion–Israel Institute of Technology, Haifa, in 1986, 1988, and 1997, respectively.

From 1988 to 1993, he served in the Israeli Air Force. From 1997 to 2000, he was with Hewlett-Packard Laboratories as an R&D Engineer. From 2000 to 2001, he headed the Research Division at Jigami Corporation, Israel. From 2001 to 2003, he was a Research Associate with the Computer Science Department, Stanford University, Stanford, CA (SCCM program). Since September 2003, he has been with the Department of Computer Science, The Technion, as an Assistant Professor. He works in the field of signal and image processing, specializing particularly in inverse problems, sparse representations, and over-complete transforms.

Dr. Elad received The Technion's Best Lecturer Award four times (1999, 2000, 2004, and 2005). He is also the recipient of the Guttwirth and the Wolf fellowships.

Michal Aharon received the B.Sc. and M.Sc. degees from the Department of Computer Science, The Technion–Israel Institute of Technology, Haifa, in 2001 and 2004, respectively. She is currently pursuing the Ph.D. degree, working closely with Prof. M. Elad and A. Bruckstein.

During her studies, she worked at Intel and IBM. From 1997 to 1999, she served in the Israeli Military Intelligence.

Ms. Aharon is the recipient of the Guttwirth and the Neeman fellowships.