Image Denoising Via Learned Dictionaries and Sparse representation

Michael Elad Michal Aharon
Department of Computer Science
The Technion - Israel Institute of Technology, Haifa 32000 Israel

Abstract

We address the image denoising problem, where zero-mean white and homogeneous Gaussian additive noise should be removed from a given image. The approach taken is based on sparse and redundant representations over a trained dictionary. The proposed algorithm denoises the image, while simultaneously trainining a dictionary on its (corrupted) content using the K-SVD algorithm. As the dictionary training algorithm 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, with state-of-the-art performance, equivalent and sometimes surpassing recently published leading alternative denoising methods.

1. 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, and as such, it has drawn a lot of research attention in the past 50 years or so. In this paper we 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*.

Redundant representations and sparsity have been used in the past decade successfully for the denoising problem. Indeed, at first, sparsity of the unitary wavelet coefficients has been considered, leading to the celebrated shrinkage algorithm [1, 2, 3, 4, 5, 6]. One reason to turn to redundant representations was the desire to have the shift invariance property [7]. Also, with the growing realization that regular separable 1D wavelets are inappropriate for handling images, several new tailored multi-scale and directional redundant transforms were introduced, including the Curvelet [8], Contourlet [9], Wedgelet [10], Bandlet [11], the steerable wavelet [12], and more. In parallel, the introduction of the matching pursuit [13, 14] and the basis pursuit denoising [15], 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 [16, 18, 19] 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-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 [20, 21].

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 pre-chosen set of basis functions as the Curvelet or Contourlet would do, we propose to learn the dictionary from examples. In this work we consider training the dictionary using patches from the corrupted image itself. 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 [22, 23, 24, 25, 26]. In this paper we propose the use of the K-SVD algorithm [25, 26], because of its simplicity and efficiency for this task. Also, due to its structure, the training and the denoising fuse together naturally into one coherent and iterated process, when training on the given image directly.

Since dictionary learning is limited in handling small image patches, we propose a global image prior that forces sparsity over those small patches in every location in the image (with overlaps). This aligns with a similar idea, appearing in [21], 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 [27].

When considering the available global and multi-scale alternative denoising schemes (e.g., based on Curvelet, Contourlet, and steerable wavelet), it looks like there is much to lose in working on small patches. In that respect, the image denoising work reported in [16] is of great importance. Beyond the specific novel and highly effective algorithm described in that paper, Portilla and his co-authors 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 [16].

In the next section we describe the way we use local sparsity and redundancy as ingredients in a global Bayesian objective. Section 3 then shows how training of the dictionary can become part of this denoising process. In Section 4 we show some experimental results that demonstrate the effectiveness of this algorithm.

2 Local to Global Bayesian Reconstruction

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.

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

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

is indeed very sparse, $\|\hat{\alpha}\|_0 \ll n$. The notation $\|\alpha\|_0$ stands for the count of the non-zero entries in α. 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 $\mathbf{D}\alpha \approx \mathbf{x}$ with a clear requirement to allow a bounded representation error, $\|\mathbf{D}\alpha - \mathbf{x}\|_2 \leq \epsilon$. Also, one needs to define how deep is the required sparsity, adding a requirement of the form $\|\hat{\alpha}\|_0 \leq L \ll n$, 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 $\|\hat{\alpha}\|_0$ non-zeros as a decaying function of some sort. Conisdering 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

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

where T is dictated by ϵ, σ, and L. The denoised image is thus given by $\hat{\mathbf{x}} = \mathbf{D}\hat{\alpha}$ [15, 28]. Notice that the above optimization task can be changed to be

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

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 because of its combinatorial nature, the matching and the basis pursuit algorithms can be used quite effectively to get an approximated solution [13, 14, 15]. Recent work established that those approximation techniques can be quite accurate, if the solution is sparse enough to begin with [28]. In this work we will make use of the orthonormal matching pursuit (OMP) because of its simplicity [14].

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 re-define 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 [19].

However, when using a specific fixed and small size dictionary $\mathbf{D} \in \mathcal{R}^{n \times k}$, this option does not exist. Reasons to confine the algorithm to use such a small dictionary could be: (i) When training takes place (as we will show in the next section), only small dictionaries can be composed; and furthermore, (ii) A small dictionary implies a locality of the resulting algorithms, which simplifies the overall image treatment. A heuristic approach for using such small dictionary 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 [27]. 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

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

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 parts of 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 been also practiced by Roth and Black when handling an MRF prior [21]. In our notations, 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 $\|\mathbf{D}\alpha_{ij}-\mathbf{x}_{ij}\|_2^2 \leq T$.

3 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 [27]. Indeed, we might do better by using a redundant version of the DCT[1], as practiced in [25]. Still, the question remains: can we make a better choice for $\mathbf{D}$ based on training? We now turn to discuss such option.

a close inspection of the global MAP penalty in (5) reveals a way to incorporate the dictionary learning into the denoising task. Returning to this equation, we can regard $\mathbf{D}$ as an unknown as well, and define our problem as

$$\left\{\hat{\mathbf{D}}, \hat{\alpha}_{ij}, \hat{\mathbf{X}}\right\} = \arg\min_{\hat{\mathbf{D}},\alpha_{ij},\mathbf{X}} \lambda\|\mathbf{X}-\mathbf{Y}\|_2^2 + \sum_{ij}\mu_{ij}\|\alpha_{ij}\|_0 + \sum_{ij}\|\mathbf{D}\alpha_{ij}-\mathbf{R}_{ij}\mathbf{X}\|_2^2. \quad (6)$$

This penalty has three kinds of unknowns: the sparse representations $\hat{\alpha}_{ij}$ per each location, the dictionary $\mathbf{D}$, and the overall output image $\mathbf{X}$. Instead of addressing all at once, we can apply a block-coordinate minimization algorithm that fixes two of the unknowns, and searching for the optimal third.

We start with an initialization $\mathbf{X} = \mathbf{Y}$ and a pre-chosen and fixed dictionary $\mathbf{D}$, seeking the optimal $\hat{\alpha}_{ij}$. In doing so, we get a complete decoupling of the minimization task to many smaller ones, of the form

$$\hat{\alpha}_{ij} = \arg\min_{\alpha} \mu_{ij}\|\alpha\|_0 + \|\mathbf{D}\alpha - \mathbf{x}_{ij}\|_2^2, \quad (7)$$

each handling one image patch. Solving this using the orthonormal matching pursuit [14] is easy, gathering one atom at a time, and stopping when the error $\|\mathbf{D}\alpha - \mathbf{x}_{ij}\|_2^2$ 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 those representations, we can turn to update the dictionary $\mathbf{D}$. Following the K-SVD algorithm, we consider updating the dictionary one column at a time [25, 26]. As it turns out, this update can be done optimally, leading to the need to perform a singular value decomposition (SVD) operation on residual data matrices, computed only on the image patches that use this atom (column). This way, the value of our penalty term is guaranteed to drop per an update of each dictionary atom, and along with this update, the representation coefficients change as well (see [25, 26] for more details), while keeping their sparsity structure.

Given all $\hat{\alpha}_{ij}$ and the updated dictionary, we can fix those and turn to update $\mathbf{X}$. Returning to (7), we need to solve

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{x}} \lambda\|\mathbf{X}-\mathbf{Y}\|_2^2 + \sum_{ij}\|\mathbf{D}\hat{\alpha}_{ij}-\mathbf{R}_{ij}\mathbf{X}\|_2^2. \quad (8)$$

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

$$\hat{\mathbf{X}} = \left(\lambda\mathbf{I} + \sum_{ij}\mathbf{R}_{ij}^T\mathbf{R}_{ij}\right)^{-1}\left(\lambda\mathbf{Y} + \sum_{ij}\mathbf{R}_{ij}^T\mathbf{D}\hat{\alpha}_{ij}\right). \quad (9)$$

[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.

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 (9) 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, a dictionary update stage, and an averaging of the resulting image patches. However, if minimization of (6) is our goal, then this process should be iterated. Given the updated $\mathbf{X}$, we can repeat the sparse coding stage, this time working on patches from the already denoised image. However, in doing so, we need to use an updated value for σ, as the noise level is now necessarily smaller. Since this value is unknown, in our experiments we chose to restrict the iterations to include sparse coding and dictionary update steps only. The averaging process in (9) was applied only once, as a final step in the algorithm.

In evaluating the computational complexity of the algorithm, we consider all the three stages - coefficient calculation (OMP process), dictionary update, and final averaging process. All stages can be done efficiently, requiring $O(nLS)$ operations per pixel, where n is the block dimension (64 in our experiments), L is the number of nonzero elements in each coefficient vector, and S is the number of iterations (10 in our experiments). 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.

4 Results

In this section we demonstrate the results achieved by applying the above methods on several test images. The tested images, as also the tested noise levels, are all the same ones as those used in the denoising experiments reported in [16], in order to enable a fair comparison.

Table 1 summarizes these denoising results. In this set of experiments, the dictionary used was 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. The redundant DCT dictionary was used as an initialization. This dictionary is described on the left side of Figure 2, where each of its atoms is shown as an 8×8 pixel image. We should note that a recently published work by Buades et. al. [17] proposes a highly effective denoising algorithm. Their method is based on the bilateral filter with a novel way of computing the weights based on block matching. Their approach, coined NL-Means, gives denoising results slightly superior to ours, but requires much more computational effort.

In all experiments, the denoising process included a sparse-coding of the patches using the OMP. The stopping rule was an average error passing a threshold, chosen empirically to be $\epsilon = 1.15 \cdot \sigma$. This means that our algorithm assumes the knowledge of σ – very much like assumed in [16]. When updating the dictionary, $(256 - 7)^2 = 62,001$ patches (all available patches from the 256×256 images, and every second patch from every second row in the 512×512 size images) were used. The denoised patches were averaged, as described in Equation (9), using $\lambda = 30/\sigma$ (this was found empirically to perform well and be robust to the various cases tried). The results shown correspond to 10 iterations of the $\{$sparse-coding, dictionary-update$\}$ process, followed by a computation of the output image based on (9).

As can be seen from Table 1, the results of the two methods are very close to each other in general. Averaging the results that correspond to [16] in this table for noise levels lower than[2] $\sigma = 50$, the value is 34.62 dB. A similar averaging over our results shows an average advantage of 0.24 dB over Portilla's method per experiment. For the higher noise power experiments, our approach deteriorates faster and achieves weaker results.

In order to better visualize the results and their comparison to those in [16], Figure 1 presents the denoising results as curves, for the images 'Peppers', 'House' and 'Barbara'. Notice that for these images, our approach outperforms the reported results of Portilla et al. for all noise levels lower than $\sigma = 50$.

Figure 4 further describes the behavior of the proposed denoising algorithm that trains the dictionary. Each iteration (containing the sparse coding and the dictionary update) improves the denoising results, where the initial dictionary is set to be the overcomplete DCT. A graph presenting this consistent improvement for several noise levels is presented in Figure 4. All graphs present the improvement over the first iteration, and therefore all curves start at zero. As can be seen, a gain of up to 1.1 dB is achievable. Figure 3 shows the results of the proposed algorithms for 'Barbara', and for $\sigma = 15$. Figure 2 presents the dictionary obtained by our training–denoising algorithm for this experiment.

5 Conclusions

This work have presented a simple method for image denoising, whose results have a state-of-the-art performance, equivalent and sometimes surpassing recently published leading alternatives. The proposed method is based on local operations and involves sparse decompositions of each image block under an evolving over-complete dictionary, and

[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.

$\sigma/PSNR$	Lena		Barb		Boats		Fgrpt		House		Pepers		σ_{PSNR}	
2/42.11	43.23	**43.58**	43.29	**43.67**	42.99	**43.14**	**43.05**	42.99	44.07	**44.47**	43.00	**43.33**	0.012	0.017
5/34.15	38.49	**38.60**	37.79	**38.08**	36.97	**38.08**	**36.68**	36.65	38.65	**39.37**	37.31	**37.78**	0.014	0.017
10/28.13	**35.61**	35.47	34.03	**34.42**	33.58	**33.64**	**32.45**	32.39	35.35	**35.98**	33.77	**34.28**	0.017	0.027
15/24.61	**33.90**	33.70	31.86	**32.37**	31.70	**31.73**	**30.14**	30.06	33.64	**34.32**	31.74	**32.22**	0.024	0.035
20/22.11	**32.66**	32.38	30.32	**30.83**	**30.38**	30.36	**28.60**	28.47	32.39	**33.20**	30.31	**30.82**	0.031	0.027
25/20.17	**31.69**	31.32	29.13	**29.60**	**29.37**	29.28	**27.45**	27.26	31.40	**32.15**	29.21	**29.73**	0.037	0.036
50/14.15	**28.61**	27.79	**25.48**	25.47	**26.38**	25.95	**24.16**	23.24	**28.26**	27.95	25.90	**26.13**	0.049	0.058
75/10.63	**26.84**	25.80	**23.65**	23.01	**24.79**	23.98	**22.40**	19.97	**26.41**	25.22	**24.00**	23.69	0.061	0.060
100/8.13	**25.64**	24.46	**22.61**	21.89	**23.75**	22.81	**21.22**	18.30	**25.11**	23.71	**22.66**	21.75	0.070	0.046

Table 1. Summary of the denoising PSNR results in [dB]. In each cell, the left value is Portilla et al. result [16], and the right refers to the proposed method. The last column presents the variance in the denoising results.

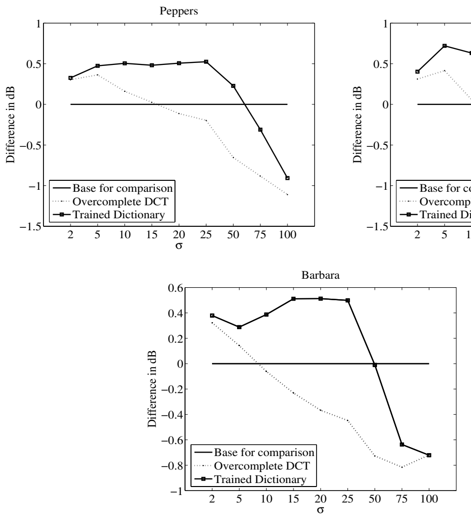

Figure 1. Comparison between Overcomplete DCT (dotted), trained dictionary (solid), and the results by [16] posed as a reference baseline.

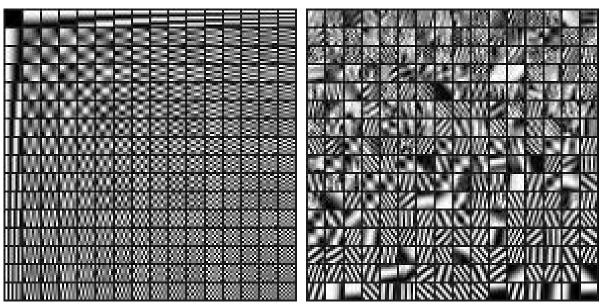

Figure 2. The overcomplete DCT dictionary (left). The trained dictionary for 'Barbara' with $\sigma = 15$, after 10 iterations (right).

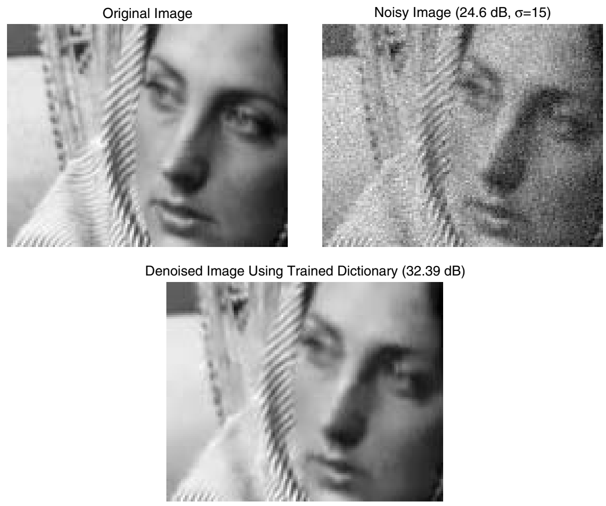

Figure 3. Zoom views of the denoising results for the image 'Barbara'.

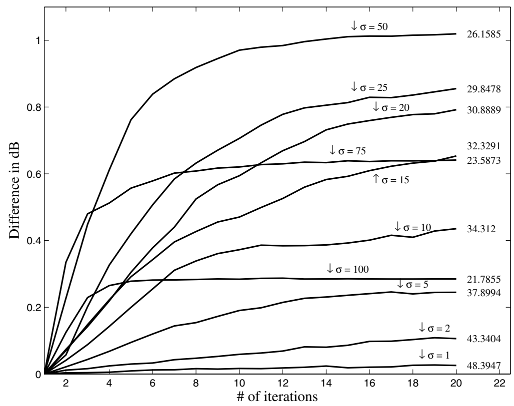

Figure 4. The improvement in the denoising results after each iteration.

a simple average calculations. The content of the dictionary is of prime importance for the denoising process – we have shown that a dictionary trained on patches of the noisy image itself performs very well.

References

[1] Donoho, D.L and Johnstone, I.M. (1994) Ideal spatial adaptation by wavelet shrinkage, *Biometrika* Vol. 81 No. 3, pp. 425–455, September.

[2] Donoho, D.L. (1995) De-noising by soft thresholding, *IEEE Transactions on Information Theory*, Vol. 41, No. 3, pp. 613–627, May.

[3] Simoncelli, E.P. and Adelson, E.H. (1996) Noise removal via Bayesian wavelet coring, Proceedings of the *International Conference on Image Processing*, Laussanne, Switzerland. September.

[4] Chambolle, A., DeVore, R.A., Lee, N.-Y., and Lucier, B.J. (1998) Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage, *IEEE Trans. Image Process.*, Vol. 7, No. 3, 319–335.

[5] Moulin, P. and Liu, J. (1999) Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors, *IEEE Transactions on Information Theory*, Vol. 45, No. 3, pp. 909–919, April.

[6] Jansen, M. (2001) *Noise Reduction by Wavelet Thresholding*, SpringerVerlag, New York.

[7] Coifman, R. and Donoho, D.L. (1995) Translation invariant de-noising. In Wavelets and Statistics, Lecture Notes in Statistics, pages 125–150, New York, 1995. Springer-Verlag.

[8] Candès, E.J. and Donoho, D.L. (2004) New tight frames of curvelets and the problem of approximating piecewise C^2 images with piecewise C^2 edges, *Comm. Pure Appl. Math.*, Vol. 57, pp. 219–266, February.

[9] Do, M.N. and Vetterli, M. (2003) Contourlets, Beyond Wavelets, G. V. Welland ed., Academic Press.

[10] Donoho, D.L. (1998) Wedgelets: Nearly minimax estimation of edges, *Annals Of Statistics*, Vol. 27, No. 3, pp. 859–897, June.

[11] Mallat, S. and LePennec, E. (2005) Sparse geometric image representation with bandelets, *IEEE Trans. on Image Processing*, Vol 14, no. 4, p. 423–438, April.

[12] Freeman, W.T. and Adelson, E.H. (1991) The design and use of steerable filters, *IEEE Pat. Anal. Mach. Intell.*, Vol. 13, no. 9, pp. 891–906, September.

[13] Mallat, S. and Zhang, Z. (1993) Matching pursuit in a time-frequency dictionary, *IEEE Transactions on Signal Processing*, Vol. 41, pp. 3397–3415.

[14] Pati, Y.C., Rezaiifar, R., and Krishnaprasad, P.S. (1993) Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition, *Proceedings of the 27 th Annual Asilomar Conference on Signals, Systems, and Computers*.

[15] Chen, S.S., Donoho, D.L. and Saunders, M.A. (2001) Atomic decomposition by basis pursuit, *SIAM Review*, Volume 43, number 1, pages 129–59.

[16] Portilla, J., Strela, V., Wainwright, M.J, and Simoncelli, E.P. (2003) Image denoising using scale mixtures of gaussians in the wavelet domain *IEEE Transactions On Image Processing*, Vol. 12, No. 11, pp. 1338–1351, November.

[17] Buades, A., Coll, B., and Morel, J. M., (2005) A non-local algorithm for image denoising, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, pp 60–65 June.

[18] Starck, J.-L., Candes, E.J., and Donoho, D.L. (2002) The curvelet transform for image denoising, *IEEE Transactions On Image Processing*, Vol. 11, No. 6, pp. 670–684, June.

[19] Matalon, B., Elad, M. and Zibulevsky, M. (2005) Improved denoising of images using modeling of the redundant contourlet transform, Proceedings of the SPIE conference wavelets, Vol. 5914, July.

[20] Zhu, S.C. and Mumford, D. (1997) Prior learning and Gibbs reaction-diffusion, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 11, pp. 1236–1250, November.

[21] Roth, S. and Black, M.J. (2005) Fields of experts: A framework for learning image priors, IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, pp. 860–867, June.

[22] Olshausen, B.A. and Field. D.J. (1997) Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, Vol. 37, pp. 311-325.

[23] Engan, K., Aase, S.O., and Hakon-Husoy, J.H. (1999) Method of optimal directions for frame design, IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 5, pp. 2443–2446.

[24] Kreutz-Delgado, K., Murray, J.F. , Rao, B.D., Engan, K., Lee, T., and Sejnowski, T.J. (2003) Dictionary learning algorithms for sparse representation. Neural Computation, Vol. 15, No. 2, pp. 349–396.

[25] Aharon, M., Elad, M., and Bruckstein, A.M. (2005) The K-SVD: an algorithm for designing of overcomplete dictionaries for sparse representation, to appear in the *IEEE Trans. On Signal Processing*.

[26] Aharon, M., Elad, M., and Bruckstein, A.M. (2005) On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them,to appear in the *Journal of Linear Algebra and Applications*.

[27] Guleryuz, O.G. (2005) Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising: Part I and Part II, to appear in the *IEEE Trans. on Image Processing*.

[28] Donoho, D.L., Elad, M., and Temlyakov, V. (2004) Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise, to appear in the IEEE Trans. Information Theory.