EFFECTIVE SEPARATION OF SPARSE AND NON-SPARSE IMAGE FEATURES FOR DENOISING

Ayan Chakrabarti and Keigo Hirakawa

Harvard University, School of Engineering and Applied Sciences
33 Oxford Street, Cambridge, MA 02138 USA
{ayanc@eecs.harvard.edu, hirakawa@stat.harvard.edu}

ABSTRACT

Over-complete representations of images such as undecimated wavelets have enjoyed immense popularity in recent years. Though they are efficient for modeling singularities and edges, natural images also consist of textures that are difficult to capture with any canonical transformation. In this work, we develop a new modeling strategy with a rigorous treatment of textured regions. Using principal components analysis as an approximate classifier for edges and textures, we partition an image into compressible and incompressible regions—with corresponding models matching their behaviors. A posterior median-based denoising method using these models is described with preliminary results that demonstrate the effectiveness of this approach.

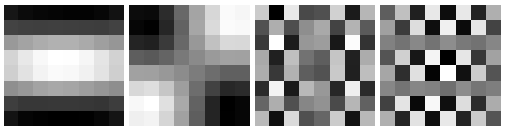
Index Terms— principal components analysis, sparsity, image denoising, image modeling, textures.

1. INTRODUCTION

Owing to the energy compaction properties, transform-based image representation provide a convenient and efficient platform for modeling image data. With the underlying assumption that the image signals live on a lower-dimensional subspace, energy compaction properties promote sparsity in the transform domain that enable image processing methods to distinguish signal from noise, and thereby help preserve image features. Redundant dictionaries, frames, and undecimated wavelets are well-documented examples of over-complete bases aimed at approximating the underlying signal with fewer basis vectors [1–5].

Explicit modeling of sparsity has been demonstrated to work well for a number of image processing applications [6–10]. For example, compressive sensing schemes reconstruct signals from its under-sampled measurements by effectively limiting the number of nontrivial transform coefficients via $\mathbf{L_1}$ minimization [7, 8]. Alternating projection methods also yield output images that are sparse in some canonical transformation [9]. Posterior median is a statistically motivated estimation techniques that takes advantage of the sparsity encoded in the form of a prior distribution on the transform coefficients [10]. The existing work in these areas have helped clarify the advantages to promoting sparsity and to incorporating it explicitly into modeling strategies.

In examining the compressibility of natural images, however, regions such as textures that do not exhibit spatial redundancies pose a challenge. Though there exist generative models used in applications such as texture synthesis [11, 12], an image feature that falls into this category is not easily generalizable because it is often a unique instance of signal energy permeating across multiple sub-bands and coefficients (i.e. incoherent with respect to the choice of basis). The sparsity-based approaches to image modeling therefore amount to treating textures as a series of small edges.



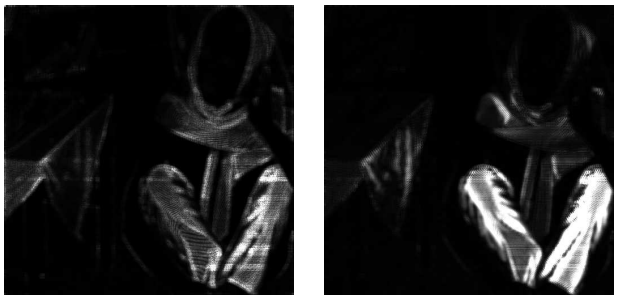
(a) $\boldsymbol{\phi}_3$ (b) $\boldsymbol{\phi}_5$ (c) $\boldsymbol{\phi}_{50}$ (d) $\boldsymbol{\phi}_{60}$

Fig. 1. PCA eigenvectors derived from training image patches (8×8). Those corresponding to large eigenvalues resemble image features such as edges and smooth regions, whereas those corresponding to small eigenvalues does not contain meaningful structures.

In light of this, we propose a statistical model of the image patch $\boldsymbol{f}$ and corresponding image denoising strategy that embody both the compressible and incompressible signal features. Given a noisy image patch $\boldsymbol{g}$, the hybridization of the sparse and non-sparse features is facilitated via principal component analysis (PCA)—trained over a large image database, PCA acts as an approximate classifier for textured and non-textured regions. Conditioned on the type of regions we identified, we devise different modeling and denoising strategies. We adopt a Bayesian statistics point of view, and model the signals in terms of the prior probability distribution of the latent variable ($p(\boldsymbol{f})$) and the likelihood of the observation conditioned on the latent variable ($p(\boldsymbol{g}|\boldsymbol{f})$). The posterior median (which minimizes the $\mathbf{L_1}$ risk) is used to estimate the image signal since it induces a thresholding rule that attenuates observed coefficients which are sufficiently small all the way to 0.

2. PCA-BASED PRIOR MODEL

In this section we consider developing a prior model for images using PCA. We restrict our attention to small individual patches of size $\sqrt{K} \times \sqrt{K}$ (where $\sqrt{K}$ is say 8). Each patch can then be thought of as a point in $\mathbb{R}^K$. We postulate that patches that appear in natural images do not occur uniformly in this space and that in fact their distribution is highly localized. Suppose we first perform PCA on patches chosen randomly from a training set of natural images—define $\{\boldsymbol{\phi}_i \in \mathbb{R}^K\}$, $i \in \{1 \ldots, K\}$ as the decorrelated orthonormal basis where the index of the eigenvectors correspond to a decreasing ordering of the magnitudes of the eigenvalues.



(a) (b)

Fig. 2. Maps of the (a) $\mathbf{L}_{0.3}$ norm in the PCA domain, and (b) Energy contained in $\mathcal{C}_I^{\perp}$ for the *Barbara* image.

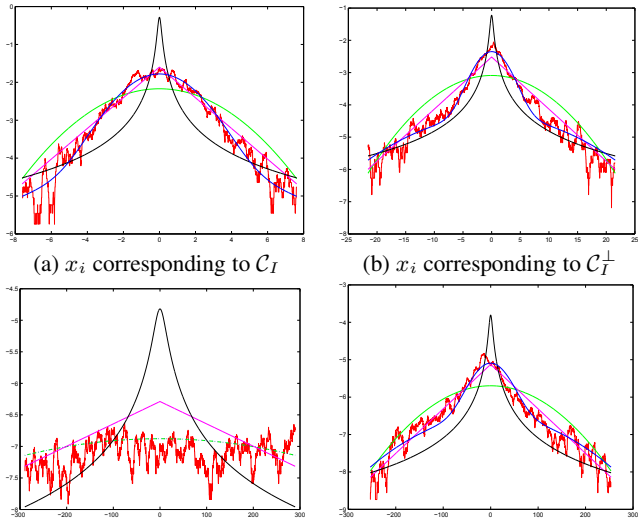
Select examples of 8×8 eigenvector patches are illustrated in Figure 1. The eigenvectors corresponding to large eigenvalues resemble edges and shapes with large features, whereas those corresponding to small eigenvalues exhibit less meaningful features. This is consistent with our presumption that spatially redundant features such as edges are compressible and sparse; and that the textures that appear in the training sets are unique instances that do not generalize to a larger class of image signals. It leads us to expect a partitioning in the PCA transform such that compressible features are contained entirely within the subspace $\mathcal{C}_I = \text{span}\{\boldsymbol{\phi}_i : i < I\}$ for some $I < K$. On the other hand, incompressible features live in $\mathcal{C}_I \oplus \mathcal{C}_I^{\perp}$ (where $\mathcal{C}_I^{\perp} = \text{span}\{\boldsymbol{\phi}_i : i \geq I\}$ is the complement of $\mathcal{C}_I$) and are likely to be incoherent with any choice of basis $\{\boldsymbol{\phi}_i \in \mathbb{R}^K\}$. It can be seen from comparing a map of the $\mathbf{L}_{0.3}$ norm (as an approximation to $\mathbf{L}_0$) of the PCA components and a map of the energy in $\mathcal{C}_I^{\perp}$ across the *Barbara* image that spatially redundant features such as smoothness and edges indeed have sparse representation in $\mathcal{C}_I$; and the regions lacking localization for a particular choice of coordinates as indicated by $\mathbf{L}_{0.3}$ coincide almost perfectly with areas of non-negligible energy concentration in $\mathcal{C}_I^{\perp}$ and with the textures in the image. We conclude that non-textured regions are sparse in the PCA domain (and in $\mathcal{C}_I$ in particular), and that the components in $\mathcal{C}_I^{\perp}$ are only required to explain textures. This suggests that statistical treatment of the components in $\mathcal{C}_I$ and $\mathcal{C}_I^{\perp}$ should be different.

Define $\boldsymbol{x} = \boldsymbol{\Phi}\boldsymbol{f}$, the PCA coefficients corresponding to the noise-free image patch $\boldsymbol{f} \in \mathbb{R}^K$, where $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1^T \ldots, \boldsymbol{\phi}_K^T]^T \in \mathbb{R}^{K \times K}$. Owing to the decorrelating properties of PCA transforms, we assume that the coefficients $\{x_i : i < I\}$ are independent and thus we model them one component at a time. As illustrated in Figure 3(a), (with the exception of x_1, the highest eigenvalue component) a typical empirical log-histogram of coefficients corresponding to vectors in $\mathcal{C}_I$ is heavy tailed, and we fit different functions to it—namely Gaussian, Laplace, t-Student, and mixture of two Gaussians. We find that the mixture of two Gaussians provides the best fits. In this paper, in order to encode the sparse nature of the coefficients explicitly, we simplify the Gaussian mixture models to a mixture of point mass (about zero) and a normally distributed variable:

$$x_i \overset{\text{i.i.d.}}{\sim} \pi_i \delta(0) + (1 - \pi_i)\mathcal{N}(0, \sigma_i^2) \qquad i < I, \qquad (1)$$

where π_i and σ_i^2 are component-dependent parameters.

The coefficient of the highest variance basis vector, x_1, corresponding to the highest eigenvalue from PCA seems to have a near-uniform distribution over a large range (See Figure 3(c)). Note that any denoising estimator assuming a uniform prior distribution



(a) x_i corresponding to $\mathcal{C}_I$ (b) x_i corresponding to $\mathcal{C}_I^{\perp}$

(c) x_1 (d) x_1 with low-pass removed

Fig. 3. The empirical log-histogram of the PCA coefficient values fitted with a t-distribution, a Gaussian distribution, a Laplace distribution and a mixture of two Gaussians(red=histogram, black=t-distribution, green=Gaussian, pink = Laplace, blue= mixture model). Near-uniform distribution in (c) has long tail after removing the low-pass component from the image, (d), yielding a better model fit.

amounts to leaving this component unchanged. In practice, it does not pose a problem because $\boldsymbol{\phi}_1$ is effectively the *mean* patch, which is typically low-pass and has the highest signal strength. Owing partly to the limited spatial support, however, the filtering induced by computing the mean of small patches is far from ideal as the attenuation at high frequencies is poor. To correct this problem, PCA is performed on patches taken from images whose low-pass components have been removed first (the actual low-pass filter may have a support larger than the size of the patches). Figure 3(d) illustrates that data now provides a better fit with the model in (1).

As illustrated in Figure 3, the empirical log-histogram of the coefficients corresponding to vectors in $\mathcal{C}_I^{\perp}$ also have a heavy tailed distribution. Coefficients in $\{x_i : i \geq I\}$, however, are not independent of each other, though they may be *uncorrelated* (to the extent that incoherence with any canonical basis implies lack of redundancy). To borrow from the strength of other channels, the coefficients $\{x_i : i \geq I\}$ are tied as:

$$[x_I, x_{I+1} \ldots, x_K]^T \sim \lambda\delta(0) + (1 - \lambda)\mathcal{N}(0, \sigma_{\perp}^2 \boldsymbol{\Sigma}), \qquad (2)$$

where λ, $\sigma_{\perp}^2$ and $\boldsymbol{\Sigma} = \text{diag}(\sigma_I^2 \ldots, \sigma_K^2)$ are parameters.

In statistics it is often accepted that the error in estimating a large number of parameters from noisy data (in the case of image denoising) offsets the benefits of having a superior, more complex model. After computing the covariance matrix of natural image patches used in PCA based on a set of randomly sampled patches from the collection of *clean* natural images, we therefore propose to estimate some of the hyper-parameters from the same training set. Assuming that the training set contains some noise of nominal variance, we train over the following model to recover $\{\pi_i, \sigma_i^2\}_{i<I}$:

$$\tilde{x_i} \sim \pi_i \mathcal{N}(0, \sigma_{ni}^2) + (1 - \pi_i)\mathcal{N}(0, \sigma_i^2 + \sigma_{ni}^2). \qquad (3)$$

The parameters are selected using the Expectation Maximization (EM) algorithm [13] to maximize the likelihood of the observed data.

To estimate the parameters for the components in $\mathcal{C}_I^\perp$, we first compute the maximal likelihood estimates of λ using EM. The other parameters, $\sigma_\perp^2$ and $\mathbf{\Sigma}$, however, are hand-picked and estimated from each noisy image itself respectively, as we expect that the texture profiles in the training sets are unique instances that does not correspond well to the noisy image at hand.

3. ESTIMATION ALGORITHM

In this section, a denoising algorithm is proposed for images corrupted by additive white Gaussian noise (AWGN) with variance σ_n^2. The noisy image is first filtered to remove the low-pass components and the PCA-based prior model is used to compute a clean estimate of the high-passed image. To this end, the high-passed image is rearranged into a set of overlapping patches $\{\boldsymbol{g}^j\}$ of size $\sqrt{K} \times \sqrt{K}$ where j is an index over patch locations. The observed patch $\boldsymbol{g}^j$ is related to the corresponding clean patch $\boldsymbol{f}^j$ as

$$\boldsymbol{g}^j = \boldsymbol{f}^j + \boldsymbol{n}^j, \boldsymbol{n}^j \sim \mathcal{N}(0, \tilde{\sigma}_n^2 \boldsymbol{I}), \tag{4}$$

where $\tilde{\sigma}_n^2$ is the variance of the noise in the high-pass component ($\tilde{\sigma}_n^2 = \sigma_n^2 \langle \boldsymbol{h}, \boldsymbol{h} \rangle$ for a high-pass filter with impulse response $\boldsymbol{h}$).

Defining $\boldsymbol{x}^j = \mathbf{\Phi} \boldsymbol{f}^j$ and $\boldsymbol{y}^j = \mathbf{\Phi} \boldsymbol{g}^j$, the posterior distribution of $\boldsymbol{x}^j$ given $\boldsymbol{y}^j$ is derived by combining the prior models in (1) and (2) with the observation model in (4). Note that as PCA generates an orthonormal basis, the noise in $\boldsymbol{y}^j$ is also AWGN with variance $\tilde{\sigma}_n^2$. The components $\{x_i^j : i < I\}$ (i.e. those in $\mathcal{C}_I$) are independent and their posterior distribution is given by

$$x_i^j | y_i^j \sim \hat{\pi}_i^j \delta(0) + (1 - \hat{\pi}_i^j) \mathcal{N}\left(\frac{\sigma_i^2 y_i^j}{\sigma_i^2 + \tilde{\sigma}_n^2}, \frac{\tilde{\sigma}_n^2 \sigma_i^2}{\tilde{\sigma}_n^2 + \sigma_i^2} \right), \tag{5}$$

where $\hat{\pi}_i^j$ is the posterior mixture weight:

$$\hat{\pi}_i^j = \frac{\pi_i \mathcal{N}(y_i^j | 0, \tilde{\sigma}_n^2)}{\pi_i \mathcal{N}(y_i^j | 0, \tilde{\sigma}_n^2) + (1 - \pi_i) \mathcal{N}(y_i^j | 0, \sigma_i^2 + \tilde{\sigma}_n^2)}.$$

The estimate $\hat{x}_i^j$ of x_i^j that minimizes the $\mathbf{L}^1$ risk in the transform domain is the median of (5) and can be shown [10] to be given by:

$$\hat{x}_i^j = \begin{cases} 0, & \text{if } \hat{\pi}_i^j \geq 0.5 \\ \text{sign}(y_i^j) \max\left(0, \frac{\sigma_i^2 |y_i^j|}{\sigma_i^2 + \tilde{\sigma}_n^2} + \sqrt{\frac{\tilde{\sigma}_n^2 \sigma_i^2}{\tilde{\sigma}_n^2 + \sigma_i^2}} \mathbf{F}^{-1}\left(\frac{0.5 - \hat{\pi}_i^j}{1 - \hat{\pi}_i^j} \right)\right), & \text{otherwise} \end{cases} \tag{6}$$

where $\mathbf{F}(x) = 1/2(1 + \text{erf}(x/\sqrt{2}))$ is the cumulative distribution function of the standard normal distribution.

Next, $\mathbf{\Sigma}$ is estimated from the observed patches $\{\boldsymbol{g}^j\}$ to fit the texture profile of the image. Specifically $\forall i \geq I$, σ_i^2 is computed as $\sigma_i^2 = \text{Var}(\{y_i^j\}_j) - \tilde{\sigma}_n^2$. These values are thresholded to ensure that they are at least four times the noise variance $\tilde{\sigma}_n^2$. The posterior distribution for the $\mathcal{C}_I^\perp$ components is similar to (5) except that the posterior mixture coefficient is pooled across all the channels and also in a small spatial neighborhood. Since the mixture coefficient here is a measure of the confidence that a particular patch contains texture, cues from neighboring patches are used by tying the mixture coefficients in a 3×3 window of neighboring patches and asserting that either all or none of them are textured. The resulting posterior distribution of $\boldsymbol{x}_I^{\perp j} = [x_I^j \ldots, x_K^j]^T$ given $\boldsymbol{y}_I^{\perp j} = [y_I^j \ldots, y_K^j]^T$ is

$$\boldsymbol{x}_I^{\perp j} | \boldsymbol{y}_I^{\perp j} \sim \hat{\Lambda}^j \delta(0) + (1 - \hat{\Lambda}^j) \mathcal{N}(\boldsymbol{\mu}^j, \tilde{\mathbf{\Sigma}}), \tag{7}$$

where $\eta(j)$ is the 3×3 spatial neighborhood of j and

$$\begin{aligned} \boldsymbol{\mu}^j &= \sigma_\perp^2 \mathbf{\Sigma} (\sigma_\perp^2 \mathbf{\Sigma} + \tilde{\sigma}_n^2 \boldsymbol{I})^{-1} \boldsymbol{y}_I^{\perp j} \\ \tilde{\mathbf{\Sigma}} &= \tilde{\sigma}_n^2 \sigma_\perp^2 \mathbf{\Sigma} (\sigma_\perp^2 \mathbf{\Sigma} + \tilde{\sigma}_n^2 \boldsymbol{I})^{-1} \\ \hat{\Lambda}^j &= \frac{\prod_{j' \in \eta(j)} \hat{\lambda}^{j'}}{\prod_{j' \in \eta(j)} \hat{\lambda}^{j'} + \prod_{j' \in \eta(j)} (1 - \hat{\lambda}^{j'})} \\ \hat{\lambda}^j &= \frac{\lambda \mathcal{N}(\boldsymbol{y}_I^{\perp j} | 0, \tilde{\sigma}_n^2 \boldsymbol{I})}{\lambda \mathcal{N}(\boldsymbol{y}_I^{\perp j} | 0, \tilde{\sigma}_n^2 \boldsymbol{I}) + (1 - \lambda) \mathcal{N}(\boldsymbol{y}_I^{\perp j} | 0, \sigma_\perp^2 \mathbf{\Sigma} + \tilde{\sigma}_n^2 \boldsymbol{I})}. \end{aligned}$$

The estimates $\{\hat{x}_i^j : i \geq I\}$ can now be computed as the medians of the marginal distributions from (7) in a similar way as in (6).

Given all the estimates $\{\hat{x}_i^j\}$, the estimate $\hat{\boldsymbol{f}}^j$ of the clean patch $\boldsymbol{f}^j$ is computed as $\hat{\boldsymbol{f}}^j = \mathbf{\Phi}^T \hat{\boldsymbol{x}}^j$; the redundant estimates due to patch overlaps are averaged to yield the denoised high-passed image.

4. EXPERIMENTAL RESULTS

We tested the proposed denoising algorithm on three images (*Cameraman*, *Lena* and *Barbara*) synthetically corrupted by AWGN. We chose the patch size to be 8×8 and the number of components in $\mathcal{C}_I^\perp$ to be 32. Training was done on 200,000 patches randomly sampled from 600 images downloaded from the website "`flickr.com`".

We compared the proposed algorithm with the wavelets-based method (using Daubechies-4 wavelets) described in [14]. Table 1 lists the quality of the denoised images for both algorithms in terms of the structured similarity index (SSIM) metric [15]. In Figure 4, we show the results of the two methods on portions of the *Barbara* image. The results for the two methods are comparable both visually and in terms of SSIM. It is interesting to note that the proposed algorithm, in general, performs better in smooth and structured regions and the wavelets-based method does better in textured regions while generating more artifacts in smooth and structured regions. This is largely because we distinguish non-textured regions from textured ones by pooling observations across channels in $\mathcal{C}_I^\perp$, and then suppress all the components in $\mathcal{C}_I^\perp$ if we detect a non-textured region. This pooling allows us to avoid the "blips" corresponding to single low-variance components that we see in the estimates of the wavelets-based algorithm. Also while the proposed method appears to be effective in separating sparse features from non-sparse ones, non-sparse features may need to be modeled more accurately for better denoising of textured regions.

5. DISCUSSION

In this work, a method has been described for modeling sparse and non-sparse image features separately using PCA on small image patches. It was shown that while sparse features like distinct shapes and edges lie almost entirely in the subspace spanned by higher-variance PCA components, the remaining components only play a role in describing textures. Independent and pooled Gaussian mixture models were used to statistically describe these two sets of components respectively. A denoising strategy was formulated based on these models with encouraging results. In future work, we plan to extend this model to separately describe the distribution of sparse components conditioned on the patch being detected as textured—possibly by using supervised learning on training patches that are labeled as textured or un-textured. We shall also explore denoising using this model under signal-dependent noise.

(a) (b) (c) (d)

Fig. 4. Comparisons on two different parts of the *Barbara* image. (a) Original image, (b) Noisy image with $\sigma_n^2 = 512$, denoised estimates using the (c) Proposed method, and (d) Algorithm described in [14].

Image	σ_n^2	Proposed method	Method in [14]
Cameraman	128	0.936078	0.941784
	512	0.878093	0.887861
	1152	0.834685	0.841675
Lena	128	0.895391	0.901787
	512	0.838272	0.853207
	1152	0.792270	0.815030
Barbara	128	0.911451	0.916306
	512	0.821288	0.840569
	1152	0.737160	0.769034

Table 1. Comparison of SSIM values for proposed method and the algorithm described in [14] on different images and noise variances.

6. ACKNOWLEDGMENTS

We would like to thank Profs. Javier Portilla and Zhou Wang for providing the code for the denoising algorithm in [14] and for computing the SSIM metric respectively. We would also like to thank Profs. Ivan Selesnick and Patrick Wolfe for useful discussions.

7. REFERENCES

[1] E. P. Simoncelli, W. T. Freeman, and E. H. Adelson, "Shiftable multi-scale transforms," *IEEE Trans. on Information Theory*, vol. 38, no. 2, pp. 587–607, March 1992.

[2] R.R. Coifman and D.L. Donoho, "Translation-invariant denoising," *Wavelets and Statistics*, vol. 103, pp. 125–150, 1995.

[3] M. Raphan and E. P. Simoncelli, "Optimal denoising in redundant bases," in *IEEE Int'l Conf. on Image Proc.*, 2007.

[4] J.-J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. on Information Theory*, vol. 50, no. 6, pp. 1341–1344, 2004.

[5] R. Gribonval and M. Nielsen, "Nonlinear Approximation with Dictionaries I. Direct Estimates," *Journal of Fourier Analysis and Applications*, vol. 10, no. 1, pp. 51–71, 2004.

[6] O.G. Guleryuz, "Nonlinear Approximation Based Image Recovery Using Adaptive Sparse Reconstructions and Iterated Denoising-Part I: Theory," *Image Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 539–554, 2006.

[7] E. J. Candes and J. Romberg, "Quantitative robust uncertainty principles and optimally sparse decompositions," *Found. of Comput. Math.*, vol. 6, pp. 227–254, 2004.

[8] D.L. Donoho, "Compressed Sensing," *IEEE Trans. on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[9] L. Mancera and J. Portilla, "L0-Norm-based Sparse Representation Through Alternate Projections," in *IEEE Int'l Conf. on Image Proc.*, 2006, pp. 2089–2092.

[10] I. M. Johnstone and B. W. Silverman, "Empirical bayes selection of wavelet thresholds," *Ann. Statist*, vol. 33, pp. 1700–1752, 2005.

[11] D. Heeger and J. Bergen, "Pyramid-based texture analysis/synthesis," in *Proc. ACM SIGGRAPH*, Aug. 1995.

[12] S. Zhu and D. Mumford, "Prior learning and Gibbs reaction-diffusion," in *IEEE Pat. Anal. Mach. Intell.*, 1997, pp. 19–11.

[13] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Stat. Soc. Series B (Method.)*, vol. 39, no. 1, pp. 1–38, 1977.

[14] J. Portilla, V. Strela, M.J. Wainwright, and E.P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. on Image Proc.*, vol. 12, no. 11, pp. 1338–1351, 2003.

[15] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Proc.*, vol. 13, no. 4, pp. 600–612, 2004.