INFORMATIVE SENSING OF NATURAL IMAGES

Hyun Sung Chang[1] *Yair Weiss*[2] *William T. Freeman*[1]

[1]MIT CSAIL
Cambridge, MA, USA

[2]Hebrew Univ. of Jerusalem
Jerusalem, Israel

ABSTRACT

The theory of compressed sensing tells a dramatic story that sparse signals can be reconstructed near-perfectly from a small number of random measurements. However, recent work has found the story to be more complicated. For example, the projections based on principal component analysis work better than random projections for some images while the reverse is true for other images. Which feature of images makes such a distinction and what is the optimal set of projections for natural images? In this paper, we attempt to answer these questions with a novel formulation of compressed sensing. In particular, we find that bandwise random projections in which more projections are allocated to low spatial frequencies are near-optimal for natural images and demonstrate using experimental results that the bandwise random projections outperform other kinds of projections in image reconstruction.

Index Terms— Compressed sensing, natural images, uncertain component analysis, informative sensing.

1. INTRODUCTION

Given a set of linear measurements $\mathbf{y} \in \mathbb{R}^p$ on a signal $\mathbf{x} \in \mathbb{R}^n$, where $\mathbf{y} = \mathbf{W}\mathbf{x}$, which choice of $\mathbf{W}$ enables the best reconstruction of $\mathbf{x}$ if $p < n$? It is well known that the optimal set of projections, which minimizes the mean squared error, can be found by the principal component analysis (PCA) when the reconstruction is linear. However, if we relax the linear recovery constraint, the optimal projections may substantially differ from the PCA projections. In this regard, compressed sensing of sparse signals [1, 2] is a spectacular demonstration of nonlinear recovery from a small number of linear projections.

The basic mathematical results in compressed sensing deal with k-sparse signals. These are signals that have at most k active (nonzero) elements, at unknown locations, in some basis. For such signals, it was shown in [1, 2], that $O(k \log n)$ generic linear measurements are sufficient to recover the signal exactly. Furthermore, the recovery can be done by a simple convex optimization or by a greedy optimization procedure [3].

However, the theory says little about what types of linear measurements are optimal for particular signals that are not ideally sparse. The basic requirement is that the measurements should be mutually incoherent with the basis in which the signal is assumed to be sparse. Random projections have most typically been used [1, 2, 4] because they prove mutually incoherent with almost any basis, but the universality of random projections does not mean that they are *universally optimal*. Elad [5] has shown that increasing the average incoherence of a measurement matrix using an iterative algorithm, can give a small increase in compressed sensing performance. For natural images, it has been found that more standard low-pass filtering (e.g. PCA projections) often gives better reconstruction results than random projections in noisy settings [6] and even in noiseless settings [7]. Lustig, Donoho, and Pauly [8] noticed that under-sampling low-pass signals less than high-pass signals can produce a better performance for real images when using a random Fourier matrix. In a similar context, Romberg [9] first takes 1,000 low-frequency DCT coefficients to get a rough sketch of the image before switching to random projections for filling in the details.

In this paper, we consider the optimized compressed sensing for natural images. Formally, for any given number of measurements, we attempt to find a set of linear projections that are maximally informative about the images based on well-known statistics of images. The optimal projections turn out to differ from PCA or random projections. We show, by experiments, that the newly found *bandwise random* projections may far outperform random projections as well as PCA projections.

2. INFORMATIVE PROJECTION: PRELIMINARY ANALYSIS

Let $\mathbf{x}$ and $\mathbf{y}$ be the original signal and a set of measurements related by $\mathbf{y} = \mathbf{W}\mathbf{x}$, where $\mathbf{W}$ is a $p \times n$ matrix ($p < n$) that consists of orthonormal row vectors. Consider the subspace $\mathcal{W}_\perp \subset \mathbb{R}^n$ which is not spanned by the row vectors of $\mathbf{W}$. It denotes the unmeasured dimension of $\mathbf{x}$. If we define $\mathbf{W}_\perp$ as an $(n-p)\times n$ matrix whose row vectors form an orthonormal basis of $\mathcal{W}_\perp$ and if we let $\mathbf{z} = \mathbf{W}_\perp \mathbf{x}$,

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{z}\end{bmatrix} = \begin{bmatrix}\mathbf{W}\mathbf{x}\\ \mathbf{W}_\perp\mathbf{x}\end{bmatrix} = \begin{bmatrix}\mathbf{W}\\ \mathbf{W}_\perp\end{bmatrix}\mathbf{x} = \mathbf{U}\mathbf{x}. \tag{1}$$

The pair of $(\mathbf{y}, \mathbf{z})$ corresponds to $\mathbf{x}$ in *rotated* basis because of the unitarity of $\mathbf{U}$. During the recovery, $\mathbf{y}$ is exactly known from the measurement outputs, while $\mathbf{z}$ should be estimated based on $\mathbf{y}$. Here, to reduce the uncertainty of $\mathbf{z}$ as much as possible, we seek to minimize the conditional entropy $h(\mathbf{z}|\mathbf{y})$. By the formula that $h(\mathbf{x}) = h(\mathbf{U}\mathbf{x}) = h(\mathbf{y}, \mathbf{z}) = h(\mathbf{y}) + h(\mathbf{z}|\mathbf{y})$, the minimization of $h(\mathbf{z}|\mathbf{y})$ is simply equivalent to the maximization of $h(\mathbf{y})$ because $h(\mathbf{x})$ is fixed. Therefore, our problem can be formally defined as

$$\mathbf{W}^* = \arg \max_{\mathbf{W}:\mathbf{W}\mathbf{W}^\intercal=\mathbf{I}} h(\mathbf{W}\mathbf{x}). \tag{2}$$

Because, in (2), we seek to maximize the uncertainty of linear projections of data, we call the optimization scheme *uncertain component analysis* (UCA) [10]. Uncertainty minimization has also been proposed recently in the context of the sequential design of compressed sensing by [11, 7]. In this context, the projections are chosen sequentially so that each projection minimizes the remaining uncertainty about the signal given the results of the previous projections.

This work was supported by a grant from Royal Dutch/Shell Group, NGA NEGI-1582-04-0004 and the Office of Naval Research MURI Grant N00014-06-1-0734.

The UCA projections are often hard to obtain in a closed form because of very complicated nature of the differential entropy. For a random vector $\mathbf{y} \in \mathbb{R}^p$ whose covariance is $\mathbf{\Sigma_y}$, its entropy $h(\mathbf{y})$ can be decomposed into

$$h(\mathbf{y}) = h(\widetilde{\mathbf{y}}) + \frac{1}{2}\ln\det(\mathbf{\Sigma_y}), \tag{3}$$

where $\widetilde{\mathbf{y}}$ is a whitened version of $\mathbf{y}$, e.g., $\widetilde{\mathbf{y}} = \mathbf{\Sigma_y}^{-\frac{1}{2}}\mathbf{y}$. Note that, in (3), $h(\widetilde{\mathbf{y}})$ is covariance-free and depends only on the shape of the probability density function (pdf) of $\widetilde{\mathbf{y}}$, while the second term solely depends on the covariance $\mathbf{\Sigma_y}$. Hence, we call $h(\widetilde{\mathbf{y}})$ the *shape term* and $\frac{1}{2}\ln\det(\mathbf{\Sigma_y})$ the *variance term*. Overall, an entropy is the sum of these terms.

In fact, each term tends to be maximized by random projections and by the PCA projections (see [12] for more details), and our objective is to accomplish a good balance between the two.

Case 1: For white data with the covariance matrix $\sigma^2\mathbf{I}$, the variance term remains constant (i.e. $p\ln\sigma$) for any choice of $\mathbf{W}$ and only the shape term plays a role. In this case, $h(\mathbf{y})$ is maximized by the projections that make the distribution as Gaussian as possible. Let $GG(\alpha)$ denote the generalized Gaussian distribution with the shape parameter α and let c_α be the shape term of its entropy (see Appendix). If the data satisfies the source separation generative model $\mathbf{x} = \mathbf{V}\mathbf{s}$ where s_i is iid, $GG(\alpha)$ and where the columns of $\mathbf{V}$ form a complete orthonormal basis, a random projection makes $p(y)$ be Gaussian when $p = 1$, as $n \to \infty$, by the central limit theorem, and then $h(y) \approx c_2 + \ln\sigma = \ln\sqrt{2\pi e}\sigma$. For $p > 1$, a set of random projections are still near-optimal, but the impact of a new (kth) random projection on the overall entropy $h(\mathbf{y})$, which we call the *capacity* of the projection and denote by $\nu(k)$, decreases with k. This is because the dependency increases with more projections. If we neglect high-order multi-information terms (beyond the pairwise dependency), the decreasing factor can be linearly approximated as [12]

$$\begin{aligned}\nu(k) &\stackrel{\text{def}}{=} E[h(y_1,\ldots,y_k) - h(y_1,\ldots,y_{k-1})] \\ &\approx c_2 - \frac{2(k-1)}{n-1}(c_2 - c_\alpha) + \ln\sigma\end{aligned} \tag{4}$$

for sufficiently large n.

Case 2: For highly non-white data, the variance term dominates the shape term, which makes the PCA projections almost optimal.

3. UNCERTAIN COMPONENTS OF NATURAL IMAGES

Interestingly, natural images have both properties from case 1 and 2. Here, we assume several well-known statistical facts about natural images: (1) Independent (or the sparsest) components of natural images form a set of Gabor-like wavelet filters in multi-resolutions [13, 14].[1] (2) If we denote each filter by $\mathbf{v}_k$ and its response to $\mathbf{x}$ by s_k, $\mathrm{Var}(s_k)$ is nearly constant for all $\mathbf{v}_k$ that are in the same band, say B_ℓ, while falling to about a fourth at the next band $B_{\ell+1}$ [15]. (3) The pdf of s_k is remarkably well modeled by $GG(\alpha)$ with $\alpha < 1$ [16, 17].

Under these assumptions, we derive a near-optimal set of projections for natural images. First, because of the highly non-whiteness

[1] In fact, the independent components are over-complete. In a specific band (resolution), each independent component corresponds to a local edge at a particular location and angle. We assume as if there were only a complete set of independent components orthogonal to each other.

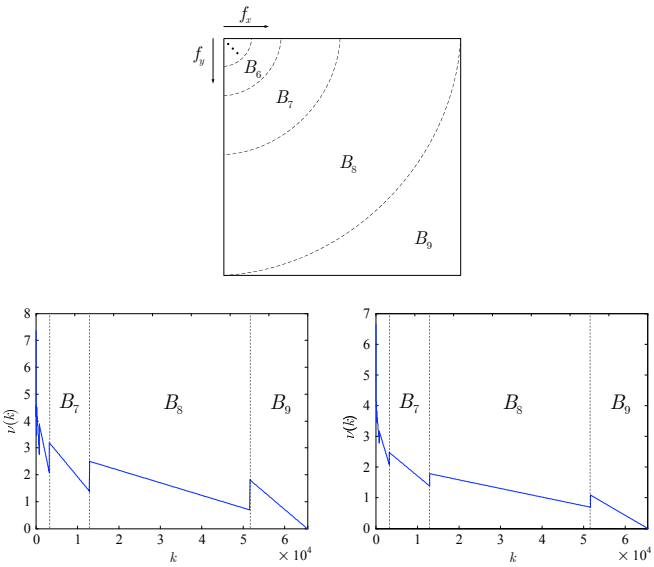

Fig. 1. Illustration of the band decomposition in spatial frequency domain (top), and capacity diagrams of bandwise random projections (bottom) for two cases $\alpha = 0.33$ (left) and $\alpha = 0.44$ (right). The assumed image size is 256×256 and then $\ell_{\max} = 9$. At the top figure, f_x and f_y denote horizontal and vertical spatial frequency, respectively.

in different bands, *cross-band* mixing is liable to decrease the overall entropy (as in case 2), and thus we seek to find a solution among the *bandwise* projections in which $\mathbf{W}$ is in the form of

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_0 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_1 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{W}_{\ell_{\max}} \end{bmatrix} \mathbf{V}^{\mathsf{T}} \tag{5}$$

where $\mathbf{V}$ is an orthonormal matrix whose columns are $\mathbf{v}_k$ and where $\mathbf{W}_\ell$ are $p_\ell \times |B_\ell|$ matrices with p_ℓ satisfying $\sum_{\ell=0}^{\ell_{\max}} p_\ell = p$. We will later discuss how to determine p_ℓ.

Next, when we sense the band-pass signals, we should use random projections (as in case 1) since they are white. Overall, the derived scheme corresponds to *bandwise random* projections. The bandwise random projections act exactly like (4) within a single band. More explicitly, for the band B_ℓ,

$$\begin{aligned}\nu(k_\ell) \approx c_2 - \frac{2(k_\ell - 1)}{|B_\ell| - 1}(c_2 - c_\alpha) + \ln(\sigma/2^\ell), \\ k_\ell = 1, 2, \ldots, |B_\ell|\end{aligned} \tag{6}$$

where k_ℓ denotes the index of each bandwise random projection taken from B_ℓ.

Finally, by the assumption of inter-band independency, we can simply concatenate $\nu(k_\ell)$ to obtain the overall capacity diagram of bandwise random projections. Fig. 1 illustrates such capacity diagrams throughout all the bands, for two example cases $\alpha = 0.33$ and $\alpha = 0.44$, as well as the band decomposition in the spatial frequency domain. Note that the overall profile varies with the value of α.

After evaluating the overall capacity $\nu(k)$, we should arrange them in the decreasing order and pick the first p projections for the

optimal choice. The optimal set of projections depends on α. Note that if the optimal number of random projections from a band p_ℓ is equal to the size of the band $|B_\ell|$ then taking p_ℓ random projections is equivalent to simply taking all the wavelet coefficients (or PCA coefficients) in that band. As Fig. 1 shows, for $\alpha = 0.44$ this happens with a small number of projections. Thus in this case, PCA is the most informative projection. However, as the number of projections increases, it is better to take random projections from different bands, while allocating more random projections to the low spatial frequencies.

4. EXPERIMENTS

In this section, we apply the UCA scheme (i.e. bandwise random projections) to natural images and make comparisons against PCA and random projections in terms of signal reconstruction performance. For the implementation, we conduct the band decomposition as we have shown in Fig. 1 but without explicit use of Gabor-like filters. Instead, we consider the DCT coefficients in the spatial frequencies between

$$\frac{2^\ell}{4\sqrt{n}} f_s \le \sqrt{f_x^2 + f_y^2} < \frac{2^\ell}{2\sqrt{n}} f_s, \qquad (7)$$

where f_s denotes the image sampling frequency in both directions. Because each DCT kernel in B_ℓ represents some harmonic (non-random) mixing of the Gabor-like wavelets in that band, the band separation in DCT domain as in (7) and subsequent band-by-band random mixing of the DCT coefficients effectively implement the proposed bandwise random projections. To carry out the random mixing, we use a set of noiselets [9, 18], binary-valued random matrix, for the efficient computer simulation.

The image recovery is based on Romberg's implementation [9], where we find the estimate $\widehat{\mathbf{x}}^*$ of the latent image by minimizing the total variation (TV), i.e.,

$$\widehat{\mathbf{x}}^* = \arg\min_{\widehat{\mathbf{x}}} \sum_{i,j} \left| \nabla \widehat{\mathbf{X}}_{ij} \right|, \quad \text{subject to} \quad \mathbf{y} = \mathbf{W}\widehat{\mathbf{x}} \qquad (8)$$

where $\widehat{\mathbf{X}}$ is the matrix representation of $\widehat{\mathbf{x}}$. The TV minimization is known to perform better than the L_1-norm minimization on the sparse basis (i.e. wavelets), avoiding high-frequency artifacts [9].

For the experiments, we used ten 256×256 images, which are from Berkeley Database [19] and shown in Fig. 2, and compared the performance, in terms of peak-signal-to-noise ratio (PSNR), of the following four projection schemes: low-frequency DCT in zig-zag order,[2] pure random projections, Romberg's method [9], and bandwise random projections. As we mentioned before, Romberg takes 1,000 low-frequency DCT coefficients and then switches to random projections.

Fig. 3 shows the reconstruction performance as a function of p, the number of projections, for two cases ($\mathcal{I}_1$ and $\mathcal{I}_7$). For the full range of p, Romberg's method (green) is better than pure random projections (cyan), which is not surprising. However, if we compare the DCT projections (blue) and Romberg's method (green), their relative performance changes completely, depending on the source image. Indeed, the two images turn out to have very different characteristics in terms of their sparsity. The generalized Gaussian parameters, roughly estimated from their wavelet coefficients, were $\alpha \approx 0.33$ for $\mathcal{I}_1$ and $\alpha \approx 0.44$ for $\mathcal{I}_7$, as their capacity diagrams are shown in Fig. 1.

[2]The DCT kernels are known to well approximate the principal components of natural images.

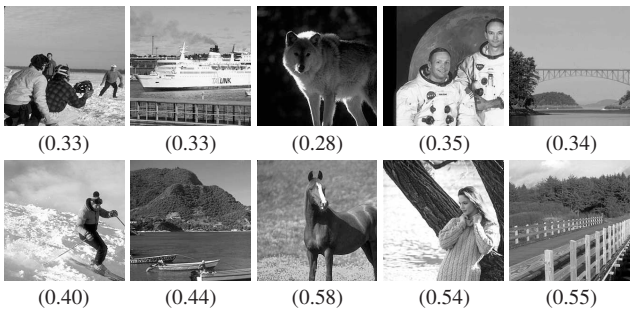

Fig. 2. Ten test images from Berkeley Segmentation Database [19], each cropped to 256×256. They are labeled $\mathcal{I}_1$–$\mathcal{I}_{10}$ from left to right, top to bottom. Each parenthesized number denotes the sparsity (GG shape parameter estimated from the wavelet coefficients) of the above image.

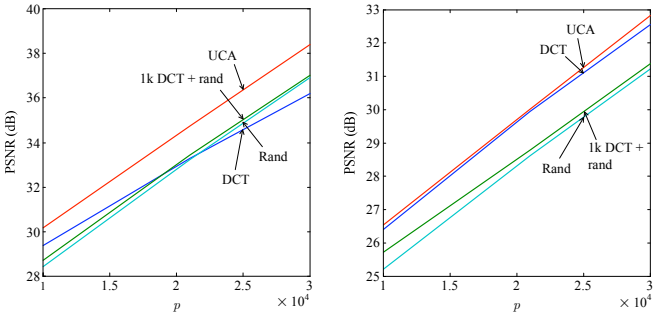

Fig. 3. Image recovery results for $\mathcal{I}_1$ (left) and for $\mathcal{I}_7$ (right). The compared schemes are the DCT projections (blue), pure random projections (cyan), Romberg's method (green), and the UCA projections (red). For decoding, the TV minimization has been used.

The image $\mathcal{I}_1$ is quite sparse. A moderate number of random projections could capture evenly all spatial frequency contents. Meanwhile, the DCT projections use up all budget for the low-frequency contents, which could be sensed with even fewer sensors, while missing high-frequency details. On the other hand, $\mathcal{I}_7$ is not that sparse. Referring to Fig. 1, we must devote almost all available budget to low-frequency bands. Otherwise, even the low-resolution version is hard to recover.

The UCA projections (red) consistently outperform Romberg's method (green) as well as the DCT projections (blue). For $\mathcal{I}_7$, the DCT projection is quite comparable to the UCA projection. This is because the DCT projection is nearly optimal for such a dense image with $\alpha \approx 0.44$.

Note that the UCA scheme uses different sets of projections depending on the sparsity of the source image. In case that the sparsity of the source image is unknown, we might have to use a value learned in advance, from a large collection of natural images. Then, we may achieve near-optimal performance in average sense, not in every case. Using a common set of bandwise random projections tuned specifically for $\alpha \approx 0.41$ (mean value) on the entire set of test images, we obtained the results as shown in Table 1. For this experiment, we used $p = 21,000$. Note that this set of projections performs best for most images (1dB better than the other projections on average) while worse than the DCT projections for the last two images ($\mathcal{I}_9$, $\mathcal{I}_{10}$). As aforementioned, the DCT projections are nearly

Table 1. The PSNR performance of image reconstruction results with $p = 21,000$. †For the UCA scheme, the same set of bandwise random projections tuned for $\alpha \approx 0.41$ has been commonly used.

Method	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	$\mathcal{I}_4$	$\mathcal{I}_5$
DCT	33.27	27.35	37.31	32.86	36.41
Rand	33.23	26.54	39.72	33.05	37.69
1k DCT + rand	33.42	26.74	39.92	33.30	37.88
UCA†	34.68	28.10	40.70	34.55	38.70

Method	$\mathcal{I}_6$	$\mathcal{I}_7$	$\mathcal{I}_8$	$\mathcal{I}_9$	$\mathcal{I}_{10}$
DCT	30.15	29.94	33.72	30.26	30.26
Rand	29.58	28.61	32.06	27.88	27.99
1k DCT + rand	29.75	28.79	32.29	28.12	28.20
UCA†	31.20	30.01	33.78	29.83	29.56

optimal for fairly dense images $\mathcal{I}_7$–$\mathcal{I}_{10}$. If we tune the bandwise random projections for $\alpha \approx 0.5$, the UCA projections give similar performance for the two images as the DCT projections.

In certain applications, it may be allowed to sense a few hundred wavelet coefficients so that we can estimate the sparsity before we do tens of thousands of projections.

5. CONCLUSION

If we are allowed to take a small number of linear projections of signals in a dataset and then use the projections plus prior knowledge of the dataset to recover the signals, what are the best projections to use? We have shown that these projections should minimize the uncertainty of a signal given its projections, or equivalently maximize the uncertainty possessed by the projections.

For natural images, we have derived a set of near-optimal projections. They are bandwise random, but allocate more sensors to the low spatial frequencies and the exact number of sensors in each band depends on the sparsity.

6. APPENDIX: GENERALIZED GAUSSIAN

A random variable x is said to be $GG(\alpha)$ if its log pdf is given by $\ln p(x) = \beta - \gamma |x - \mu|^\alpha$, for some β, γ, μ. Laplacian ($\alpha = 1$) and Gaussian ($\alpha = 2$) belong to this family and the distribution becomes sparser as $\alpha \to 0$. The shape term of $GG(\alpha)$ is computed to

$$c_\alpha = \frac{1}{2} \ln \left(\frac{4}{\alpha^2} \frac{\Gamma^3\left(\frac{1}{\alpha}\right)}{\Gamma\left(\frac{3}{\alpha}\right)} \right) + \frac{1}{\alpha} \quad \text{(nats)}, \tag{9}$$

as drawn in Fig 4.

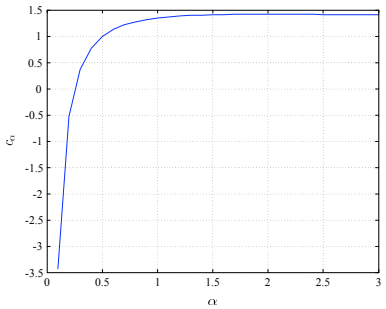

Fig. 4. Unit-variance entropy (shape term) of $GG(\alpha)$ for various values of α.

7. REFERENCES

[1] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[2] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[3] M. F. Duarte, M. B. Wakin, D. Baron, and R. G. Baraniuk, "Universal distributed sensing via random projections," in *Proc. of International Conference on Information processing in Sensor Networks*, Apr. 2006, pp. 177–185.

[4] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 83–91, Mar. 2008.

[5] M. Elad, "Optimized projections for compressed sensing," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5695–5702, Dec. 2007.

[6] J. Haupt and R. Nowak, "Compressive sampling vs. conventional imaging," in *Proc. IEEE ICIP*, Oct. 2006, pp. 1269–1272.

[7] M. W. Seeger and H. Nickisch, "Compressed sensing and Bayesian experimental design," in *Proc. ICML*, June 2008, pp. 912–919.

[8] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn. Reson. Med.*, vol. 58, no. 6, pp. 1182–1195, Dec. 2007.

[9] J. Romberg, "Imaging via compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 14–20, Mar. 2008.

[10] Y. Weiss, H. S. Chang, and W. T. Freeman, "Learning compressed sensing," in *Proc. Allerton Conf. on Communication, Control, and Computing*, Sept. 2007.

[11] R. M. Castro, J. Haupt, R. Nowak, and G. M. Raz, "Finding needles in noisy haystacks," in *Proc. IEEE ICASSP*, Mar. 2008, pp. 5133–5136.

[12] H. S. Chang, Y. Weiss, and W. T. Freeman, "Informative sensing," arXiv:0901.4275 [cs:IT], Jan. 2009.

[13] A. J. Bell and T. J. Sejnowski, "Edges are the independent components of natural scenes," in *NIPS*, 1997, pp. 831–837.

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

[15] A. van der Schaaf and J. H. van Hateren, "Modelling the power spectra of natural images: statistics and information," *Vision Research*, vol. 36, no. 17, pp. 2759–2770, 1996.

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

[17] Y. Weiss and W. T. Freeman, "What makes a good model of natural images?" in *Proc. IEEE CVPR*, June 2007.

[18] E. J. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Prob.*, vol. 23, no. 3, pp. 969–986, June 2007.

[19] D. R. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE ICCV*, July 2001, pp. 416–423.