A compressed sensing model of crowding in peripheral vision

Jens Hocke[a] and Michael Dorr[b] and Erhardt Barth[a]

[a]Institute for Neuro- and Bioinformatics, University of Lübeck, Ratzeburger Allee 160, D-23538 Lübeck, Germany
[b]Schepens Eye Research Institute, Harvard Medical School

ABSTRACT

We here model peripheral vision in a compressed sensing framework as a strategy of optimally guessing what stimulus corresponds to a sparsely encoded peripheral representation, and find that typical letter-crowding effects naturally arise from this strategy. The model is simple as it consists of only two convergence stages. We apply the model to the problem of crowding effects in reading. First, we show a few instructive examples of letter images that were reconstructed from encodings with different convergence rates. Then, we present an initial analysis of how the choice of model parameters affects the distortion of isolated and flanked letters.

Keywords: peripheral vision, crowding, compressed sensing, sparseness

1. INTRODUCTION

The human visual system samples only a tiny part of the visual field with high resolution, so that peripheral vision is not only blurred, but also heavily subsampled. Nevertheless, we experience the illusion of high-fidelity perception in the periphery despite objectively impaired recognition performance.

Reduced acuity is only one factor that limits peripheral vision. More importantly, peripheral vision suffers from crowding, such that objects may be recognized when presented in isolation, but not when presented near flanker objects.[1] Although crowded objects cannot be recognized and appear jumbled, they can be detected and are perceived as having normal contrast. A recent review[2] lists the following further properties of crowding: i) crowding depends on the spacing between target and flankers, and the critical spacing range increases with eccentricity; ii) crowding is anisotropic; iii) crowding is asymmetric.

Even though most demonstrations show spatial crowding only, crowding is not limited to the spatial domain. Temporal flankers, i.e., objects shown in the same location as the stimulus before and after stimulus presentation, can also impair identification.

A large body of literature has described a variety of approaches to model crowding, where the main categories comprise pooling, substitution, and masking models – for a recent review, see Levi.[1] Pooling models average measured features over large pooling areas, similar to receptive fields, and only the output after pooling is available for recognition. In substitution models, individual features can be detected, but lack information on precise location and object identity, so that features of the target stimulus may be substituted by features of the flankers. Finally, masking models suggest that strong features of flankers suppress features of the target. While all these approaches can capture some quantitative aspects of crowding, they fail to explain why the subjective experience of the visual periphery is relatively undegraded.

We here propose an alternative pooling model for retinal vision. It is based on concepts from compressed sensing,[3,4] a framework that allows to sample signals and images at a significantly lower rate than predicted by the Shannon-Nyquist theorem. This low sampling rate is achieved using random samples and requires that the signal is sparse, like natural images are. Particularly good compression rates in technical applications of compressed sensing were obtained with Poisson-disk instead of truly random sampling;[5] for our model, such use of a Poisson-disk sampling mosaic can also be motivated biologically because the retinal photoreceptor mosaic follows a similar pattern.[6] Nevertheless, our model in its current form is not a detailed model that would

Further author information: (Send correspondence to Jens Hocke)
Jens Hocke: E-mail: hocke@inb.uni-luebeck.de

account for all physiological data, but simplifies retinal processing to the following stages: (i) first convergence and sampling stage: lateral connections in the retina provide local integration (modeled as Gaussian blur plus Poisson-disc sampling) with a degree of blur and sampling that both depend on eccentricity; this creates a blurred and sub-sampled image; (ii) second convergence and sampling stage: retinal output is obtained by compressed sensing, i.e., a random sampling of the blurred image with a convergence rate (ratio of input and output neurons) that, again, depends on eccentricity.

The first convergence stage models the well-known fact that image blur is highly eccentricity-dependent, and varies up to 1000-fold in the human retina.[7,8] In the fovea the blur is small because of the tight packing of photoreceptors (dense sampling) and the low level of neural integration (one cone signal is processed by up to several ganglion cells); the periphery, however, shows low cone density and high neural integration and convergence,[8] and therefore a larger blur combined with sparser sampling.

In the second convergence stage, the input neurons are randomly connected to output neurons. There are far fewer output neurons than input neurons, so each output neuron receives input from several input neurons and input neurons may project to multiple output neurons. The randomized collection of information by the output neurons corresponds to what the sensing matrix does in compressed sensing (see below) and allows to transmit a signal with reduced bandwidth. In the visual system, this is useful because images need to be transmitted from the retina to visual cortex through the optic nerve, which acts as a bandwidth-limited bottleneck.

In technical systems, signals must be decoded after transmission. Compressed sensing uses the sparseness of natural signals to reconstruct the original image; mathematically, the reconstruction is equivalent to the solution of an underdetermined system of equations. The brain, on the other hand, does not need to reconstruct the original image. However, to obtain an intuitive measure of how much information was encoded by the relatively few output neurons (in other words, how much information was lost during transmission), we can use standard compressed sensing tools to visualize the reconstructed image.

This model can be considered a pooling model since the incoming features are combined (pooled). However, in contrast to alternative pooling models, the features are combined randomly; the similarities of features, such as similar orientation,[9] are not taken into account.

Our model is also related to the work by Coulter et al.,[10] who proposed an adaptive version of compressed sensing as a model of visual processing in the primary visual cortex (V1). This was motivated by sparse coding models that learn the response properties of simple cells in V1. Instead of one-to-one connections, simple cells are randomly connected to the input neurons and therefore adaptive compressed sensing allows for a more realistic model of V1 connectivity. These authors could experimentally show that the same response properties as in standard sparse coding can be learned with a random compressed-sensing network. Furthermore, they proposed that such a model of random sampling could be used between cortical areas to avoid the need of a one-to-one connectivity.

The compressed-sensing model presented here is a model of the retina instead of V1 and, accordingly, adds a blurring and subsampling stage to the adaptive compressed sensing model. The blur and the subsampling are motivated by the properties of the foveal and peripheral retina. In contrast to Coulter et al., however, we did not specifically learn a sparse basis for natural images. Instead, we used the Dual Tree Complex Wavelet Transform (DT-CWT) for practical reasons: by using the computationally efficient DT-CWT, it was possible to run simulations faster and on larger image patches.

2. MATHEMATICAL MODEL

To understand our model of early vision we need to have a basic understanding of compressed sensing (CS).[3,4] It is a sampling scheme that allows images to be sampled at a significantly lower rate than predicted by the Shannon-Nyquist theorem. To achieve this low sampling rate, the measurements are acquired by correlating a signal $\mathbf{x}$ with the sensing waveforms θ_k:

$$y_k = \langle \mathbf{x}, \theta_k \rangle, \; k = 1, 2, \ldots, m. \tag{1}$$

The entries of the sensing waveforms θ_k, used for CS, are usually generated with a certain degree of randomness[3,11] to ensure global measurements of the signal and maximum difference between the sensing waveforms.

The sampling with several sensing waveforms can be expressed by a measurement matrix Θ with m sensing waveforms θ_k as rows:

$$\mathbf{y} = \Theta\mathbf{x}. \tag{2}$$

If $\mathbf{x}$ is then represented using a sparse basis Ψ we obtain

$$\mathbf{y} = \Theta\Psi\mathbf{a}, \tag{3}$$

and one can find a good solution for the system of equations defined by Equation 3 although the system is underdetermined (the dimension of $\mathbf{y}$ is lower than that of $\mathbf{x}$). Given signals that are s-sparse, i.e. $\mathbf{a}$ has s non-zero elements, in the basis Ψ, one can use the sparseness constraint

$$\arg\min_{\mathbf{a}} ||\mathbf{y} - \Theta\Psi\mathbf{a}||^2 \text{ s.t. } ||\mathbf{a}||_0 < s \tag{4}$$

that defines an optimization problem, which can be approximately solved by using the method of compressive sampling matching pursuit (CoSaMP).[12]

In our model we start with an image $\mathbf{x}$ that corresponds to the image sampled by the receptors. In the first convergence stage (see above) the image is blurred and sub-sampled. The blur is Gaussian and is followed by a Poisson-disk sampling. The blur can be described by the convolution matrix C and the resulting image is denoted $\mathbf{x}_{\text{blur}}$. This blurred image is re-sampled to yield $\mathbf{x}_{\text{sampled}}$:

$$\mathbf{x}_{\text{sampled}} = \Phi\mathbf{x}_{\text{blur}}, \tag{5}$$

where $\Phi \in \mathbb{R}^{n\times m}$ is a sampling matrix that subsamples $\mathbf{x}_{\text{blur}}$ ($n < m$). The Poisson-disk distribution is generated using the algorithm of Bridson[13] and the resulting sampling patterns are illustrated in Figure 1. The convergence rate defined by blurring and sub-sampling is considered to increase smoothly from the fovea to the periphery. For the purpose of this paper, however, we simplified this aspect and considered only block-wise constant convergence rates, i.e. the convergence rates have been varied to simulate foveal and peripheral vision but the images have been sampled uniformly across the test letters that we considered.

The samples defined by Equation 5 are then considered to be inputs to a network of neurons such that the output neurons are randomly connected to the input neurons with random weights. The activation of the output neurons can then be computed using the connection matrix Θ:

$$\mathbf{x}_{\text{output}} = \Theta\mathbf{x}_{\text{sampled}}. \tag{6}$$

Every row in the connection matrix represents an output neuron. For every input neuron that is connected to an output neuron there is a non-zero entry in the corresponding column. The values in the matrix represent the weights of the connections.

Convergence is modeled by allowing for fewer output neurons than input neurons. The convergence rate (ratio of input and output neurons) varies with eccentricity. As for the first convergence stage, we used different

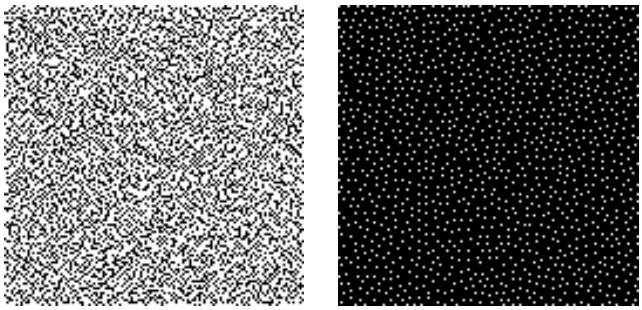

Figure 1. A foveal (left) and a peripheral (right) distribution of the sampling scheme used in the first stage. The Poisson disk radius is 1 pixel for foveal sampling and 3 pixels for peripheral sampling.

convergence rates but the convergence rate for a particular letter image did not vary across the image (for computational reasons only).

Due to convergence, the problem of reconstructing the input pattern from the output neurons is underdetermined, similar to the reconstruction problem in compressed sensing (Equation 4). Despite the compression efficiency of CS, there will be a loss of information if the input signal contains more information than can be encoded by the output neurons. By reconstructing the original image from the activity of the output neurons we estimate the amount of information delivered at the output. To reconstruct the original image, the signal at the output neurons is modeled as a linear superposition in a sparse basis Ψ:

$$\mathbf{x}_{\text{output}} = \Theta\Phi\text{C}\Psi\mathbf{a} \tag{7}$$

where $\mathbf{x}_{\text{blur}} = \text{C}\mathbf{x}$ and $\mathbf{x} = \Psi\mathbf{a}$. The Dual Tree Complex Wavelet Transform is used as basis, which is biologically plausible in that its basis elements are oriented Gabor-like functions similar to simple cells in primary visual cortex.[14,15] The basis is sparse in the sense that relatively few coefficients are needed to describe natural images in this basis. The CoSaMP algorithm is then applied to solve the following optimization problem:

$$\arg\min_{\mathbf{a}} ||\mathbf{x}_{\text{output}} - \Theta\Phi\text{C}\Psi\mathbf{a}||^2 \text{ s.t. } ||\mathbf{a}||_0 < s. \tag{8}$$

From the recovered coefficients a the image $\hat{x}$ can be then reconstructed. This image visualizes the information, which is still represented by the (rather few) samples in x_{output}.

3. EXPERIMENTS

An everyday activity that is clearly affected by crowding is reading: at a certain eccentricity a single letter may be readable, but when the letter is flanked by other letters (i.e., embedded in a word), the same letter may not be readable. Once foveated, however, even flanked letter can be easily identified.

To simulate this phenomenon qualitatively, 10 images each of single letters and of letters flanked by two random letters were generated. The two letters were placed to the left and to the right of the central letter. The images had a size of 128 by 128 pixels and bright letters (about 30 pixels high) were placed in the center on a dark background. Foveal and peripheral representations of these letters were simulated by using our model. For the fovea only a small Gaussian blur of variance 2 was used and the radius for the Poisson disk sampling was just 1 pixel, which resulted in roughly 10800 samples for the 16384 pixels of the image. In the periphery, the images were blurred with a Gaussian of variance 4 pixels, and the Poisson disk had a radius of 3 pixels, which resulted in about 1750 samples. Convergence in the second model stage, i.e., the number of output neurons, was varied from 1500 to 50, resulting in overall convergence rates of 11 to 328.

To estimate the amount of information represented by the final samples, the input images were reconstructed as described above. The DT-CWT was used with 5 levels of evaluation depth. In Figure 2 example reconstructions are shown. To compare the input and output images, the mean squared error was used as a difference measure. The mean squared error was computed only within a bounding box around the central letter to avoid measuring errors due to the flanking letters.

We tested the reconstruction quality depending on the number of output neurons, and for the above selected values of the Gaussian blur and Poisson-disc sampling. The errors were measured for all images and averaged over 10 images.

4. RESULTS

In Figure 2 one can see that the foveal output neurons have sufficient information encoded to allow for a good reconstruction for both the un-flanked and the flanked letter image. The overall quality of the reconstruction from the peripheral output neurons is of course worse, but the letter 'X' is still recognizable if it is not flanked by other letters. Note that the reconstruction is based on only 1.5% of the input pixels in case of a convergence rate of 66 in stage 2 (250 output neurons). If the letter is flanked by other letters the reconstructed images exhibit typical crowding artifacts in the sense that the letter can be located, but its features are jumbled and

recognition is impossible. However, if the convergence rate in stage 2 is decreased to 27 (600 output neurons), also the flanked letter 'X' becomes recognizeable.

Figure 3 shows mean squared error (MSE) as a function of the overall convergence rate. The parameters of the first convergence stage have been kept fixed at the same two foveal and the peripheral values that were used to create the images in Figure 2 – see previous section. Only the number of output neurons in the second convergence stage has been varied. First note that the errors increase with increasing convergence rate and that the two error curves (single letter vs. flanked letter) diverge. This shows us that the crowding effect in our model is due to a limited encoding capacity of the output neurons. However, the first convergence stage does also play a role since the results for the foveal and the peripheral simulation differ. It seems particularly interesting that the critical convergence rate (where the errors start rising) differ for the four curves. Especially in the right plot (peripheral first stage), it becomes evident that a convergence rate optimized for single letters is not optimal for letter with flankers. In the left plot (foveal first stage) the difference between the two optima is small. Note also that the reconstruction errors are in all cases higher for the flanked letter. For both settings the errors of flanked and un-flanked images do not reach the same level. The difference between the errors becomes smaller for the foveal model. Probably the residual errors when many output neurons are used are not due to a limited coding capacity, but due to the under-sampled deconvolution problem. This deconvolution problem sets a limit to the best possible reconstruction. With a full sampling of a sharp image by the receptors and the use of one fourth of the number of pixels as output neurons, a perfect reconstruction is possible for any number of flankers, since that is a standard compressed sensing problem.

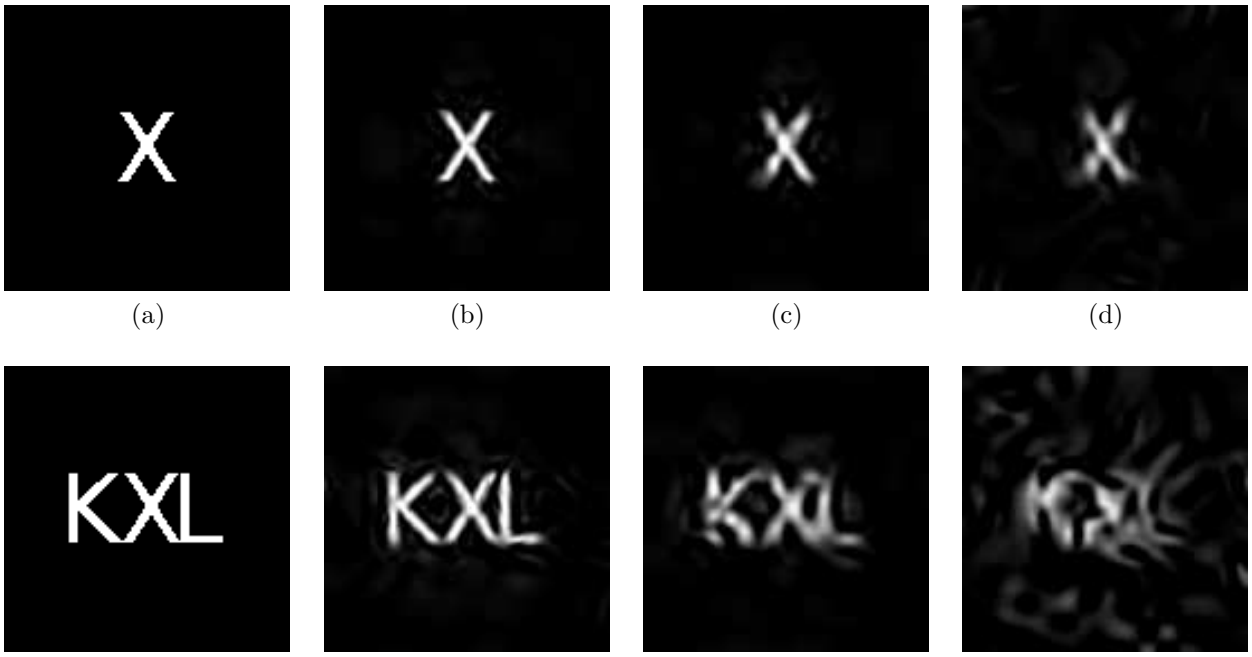

Figure 2. The letter X is shown in isolation (a) and flanked by other letters (e). All other images show reconstructed versions of these input images. The letters can still be easily recognized after reconstruction from a high-bandwidth foveal representation (b, f). These two images correspond to the two arrows in the left plot of Figure 3; the convergence rate is 13, i.e. the second stage has 1250 output neurons. The images on the right (c, d, g, h) are reconstructed from the peripheral first stage (see text) with two different convergence rates in the second stage. The four images correspond to the four arrows in the right plot of Figure 3. The two convergence rates are 27 and 66, i.e., the second stage had 600 and 250 output neurons, respectively. These four images illustrate the available information at convergence rates where the flanked letter turns unrecognizable while the un-flanked letter can still be read.

5. DISCUSSION

It is well known that natural images exhibit redundancies. As a consequence, natural images can be compressed. A more specific insight from the theory of sparse coding is that natural images can be encoded with few coefficients in a proper basis. The framework of compressed sensing provides further insight by showing that if a signal can be encoded in a sparse way, the number of samples one needs to sample the image is much lower than indicated by the classical Shannon/Nyquist rate. To explore the consequences of this for retinal coding, we proposed a simple model of low-level vision that uses compressed sensing for the coding of input signals. The model consists of two convergence stages. The convergence rates of these two stages, i.e., the blur in the first stage and the number of output neurons in the second stage, may reduce the amount of information that can be represented. As we move towards representations with loss of information, i.e., we move towards the periphery, we find a clear crowding effect in the sense that the reconstruction errors for letters images are significantly higher when the letters are flanked by other letters.

We have not provided a full analysis of the model. Instead, we have fixed the parameters of the first stage to two (foveal and peripheral) values and have varied the convergence rate of the second stage for those two values. We have found that for both locations in the first stage, the reconstruction error increases with the convergence rate of the second stage. The errors are lower at the foveal location of the first stage and they are lower for the un-flanked letter than for the flanked letter. A more interesting result is that the two error curves (flanked and un-flanked) are qualitatively different at the two locations (foveal and peripheral in the first stage). To optimize the representation, one would choose a second-stage convergence rate as the rate at which the error starts rising. At the foveal location it seems possible to choose a convergence rate that is optimal for both the un-flanked and flanked letters, see arrows in the left plot of Figure 3. At the peripheral location, however, it seems that one can go for a higher convergence rate for the un-flanked letter than one could for the flanked letters. Therefore, if the peripheral system had been optimized for the recognition of un-flanked letters, flanked letters may already be unreadable. Our reconstructed images in Figure 2 can give some guidance as to where the thresholds may lie.

We believe that out model is simple compared to alternative pooling models. We avoid, for example, assumptions such as the combination of similar features in the models of Balas et al.[16] and Freeman et al.[9] Although we here modeled crowding at the retinal level, we believe that there may also be crowding effects due to processing

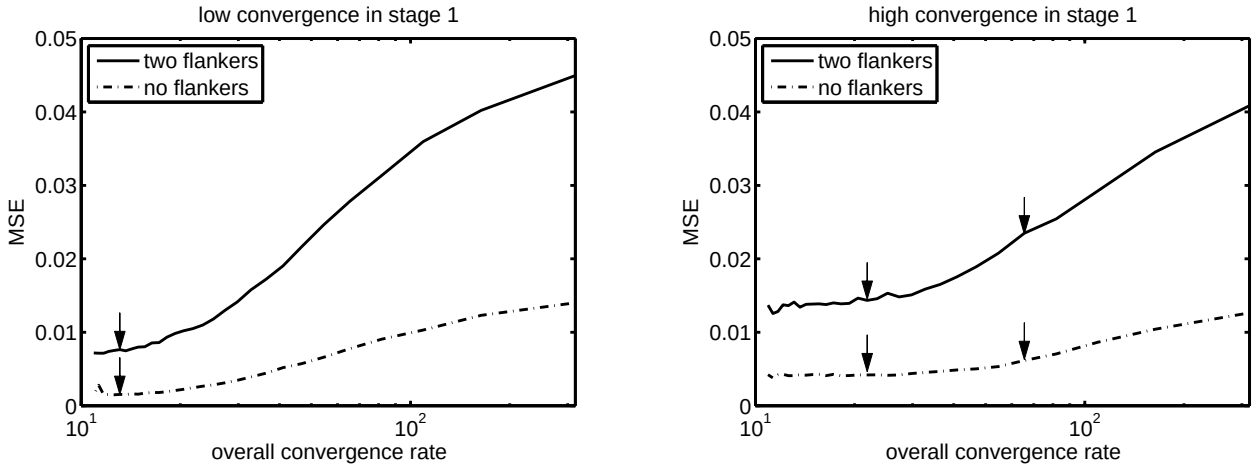

Figure 3. The graphs show the reconstruction errors for a foveal (left) and a peripheral (right) sample of the first stage and as a function of the total convergence ratio (number of input samples in x divided by the number of output samples in x_{output}), i.e., as a function of the number of output neurons in the second stage. The parameters of the first stage are those given in the text. The six arrows correspond to the six reconstructed images shown in Figure 2 and the values of the parameters are given in the caption of Figure 2. Note that, in both plots, the reconstruction error increases with the convergence rate and that the errors are lower for the un-flanked letter (dotted line) than for the flanked letter. Also note that the two error curves (flanked and un-flanked) are qualitatively different in the two plots.

in later stages. Many open questions remain and dedicated experiments need to be performed to address specific issues of the model assumptions.

In conclusion, we hope to have shown that the principles of sparse coding and compressed sensing have the potential to provide some new insights regarding visual coding in general and the crowding phenomenon in particular.

REFERENCES

[1] Levi, D., "Crowding–an essential bottleneck for object recognition: A mini-review," *Vision Research* **48**(5), 635–654 (2008).

[2] Whitney, D. and Levi, D., "Visual crowding: a fundamental limit on conscious perception and object recognition," *Trends in Cognitive Sciences* **15**(4), 160–168 (2011).

[3] Candès, E. J. and Tao, T., "Decoding by linear programming," *IEEE Transactions on Information Theory* **51**(12), 4203–4215 (2005).

[4] Donoho, D. L., "Compressed sensing," *IEEE Transactions on Information Theory* **52**(4), 1289–1306 (2006).

[5] Lustig, M., Alley, M. T., Vasanawala, S., Donoho, D. L., and Pauly, J. M., "Autocalibrating parallel imaging compressed sensing," in [*Proceedings of the 17th Annual Meeting of ISMRM*], 379 (2009).

[6] Yellott, J., "Spectral consequences of photoreceptor sampling in the rhesus retina," *Science* **221**, 382–385 (1983).

[7] Curcio, C. A., Sloan, R. R., Packer, O., Hendrickson, A. E., and Kalina, R. E., "Distribution of cones in human and monkey retina: individual variability and radial asymmetry," *Science* **236**, 579–582 (1987).

[8] Curcio, C. and Allen, K., "Topography of ganglion cells in human retina," *The Journal of Comparative Neurology* **300**(1), 5–25 (1990).

[9] Freeman, J. and Simoncelli, E. P., "Metamers of the ventral stream," *Nature Neuroscience* **14**, 1195–1201 (Aug. 2011).

[10] Coulter, W., Hillar, C., Isley, G., and Sommer, F., "Adaptive compressed sensing - A new class of self-organizing coding models for neuroscience," in [*Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*], 5494–5497, IEEE (2010).

[11] Candès, E. J. and Wakin, M. B., "An introduction to compressive sampling," *IEEE Signal Processing Magazine* , 21–30 (March 2008).

[12] Needell, D. and Tropp, J. A., "CoSaMP: iterative signal recovery from incomplete and inaccurate samples," *Communications of the ACM* **53**(12), 93–100 (2010).

[13] Bridson, R., "Fast poisson disk sampling in arbitrary dimensions," in [*ACM SIGGRAPH 2007 sketches*], 22–es, ACM (2007).

[14] Kingsbury, N., "The dual-tree complex wavelet transform: a new efficient tool for image restoration and enhancement," in [*EUSIPCO: European Signal Processing Conference*], 319–322 (1998).

[15] Selesnick, I., Baraniuk, R., and Kingsbury, N., "The dual-tree complex wavelet transform," *IEEE Signal Processing Magazine* **22**(6), 123–151 (2005).

[16] Balas, B., Nakano, L., and Rosenholtz, R., "A summary-statistic representation in peripheral vision explains visual crowding," *Journal of Vision* **9**(12), 13 (2009).