PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Learning real and complex overcomplete representations from the statistics of natural images

Bruno A. Olshausen, Charles F. Cadieu, David K. Warland

Bruno A. Olshausen, Charles F. Cadieu, David K. Warland, "Learning real and complex overcomplete representations from the statistics of natural images," Proc. SPIE 7446, Wavelets XIII, 74460S (4 September 2009); doi: 10.1117/12.825882

Event: SPIE Optical Engineering + Applications, 2009, San Diego, California, United States

SPIE.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 1/14/2018 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Learning real and complex overcomplete representations from the statistics of natural images

Bruno A. Olshausen[a], Charles F. Cadieu[b] and David K. Warland[c]

[a,b]Redwood Center for Theoretical Neuroscience, Helen Wills Neuroscience Institute and
[a]School of Optometry, University of California, Berkeley;
[c]Department of Neurobiology, Physiology and Behavior, University of California, Davis

ABSTRACT

We show how an overcomplete dictionary may be adapted to the statistics of natural images so as to provide a sparse representation of image content. When the degree of overcompleteness is low, the basis functions that emerge resemble those of Gabor wavelet transforms. As the degree of overcompleteness is increased, new families of basis functions emerge, including multiscale blobs, ridge-like functions, and gratings. When the basis functions and coefficients are allowed to be complex, they provide a description of image content in terms of local amplitude (contrast) and phase (position) of features. These complex, overcomplete transforms may be adapted to the statistics of natural movies by imposing both sparseness and temporal smoothness on the amplitudes. The basis functions that emerge form Hilbert pairs such that shifting the phase of the coefficient shifts the phase of the corresponding basis function. This type of representation is advantageous because it makes explicit the structural and dynamic content of images, which in turn allows later stages of processing to discover higher-order properties indicative of image content. We demonstrate this point by showing that it is possible to learn the higher-order structure of dynamic phase—i.e., motion—from the statistics of natural image sequences.

Keywords: sparse coding, natural images, overcomplete dictionaries, complex wavelets, motion, video

1. INTRODUCTION

Overcomplete representations have a number of desirable properties for image coding: they allow for "shiftability," such that translation or other transformations in the image result in smooth and easily predictable changes among the coefficients;[1] they provide robustness in situations where coding precision is limited;[2] and they can provide highly compact, sparse representations suitable for video compression.[3,4] One of the central questions that arises in the design of an overcomplete representation is the choice of dictionary. Ideally, one would like the elements of the dictionary to match the structures contained in images. For this reason Gabor[3] or Gaussian[4] atoms of various aspect ratios have been used to capture lines and edges in images. However, for the diverse forms of structure that occur in natural images it is difficult to know *a priori* what class of functions is most appropriate. The optimal choice of dictionary ultimately depends upon image statistics as well as task demands. Here, we focus on the contribution from image statistics.

In previous work, we showed how an overcomplete dictionary may be adapted to the statistics of the data so as to form a *sparse representation* of image content.[5,6] That is, the dictionary is tailored to the structure in natural images so that only a small fraction of atoms within the dictionary are needed to approximate any given image. Such a representation is desirable because it provides a compact encoding that makes explicit the presence of structural features within the image. Previously however we explored representations that were only modestly overcomplete (factor of two), which resulted in a Gabor-like dictionary. Our goal here is to explore higher degrees of overcompleteness, and to perform a systematic analysis of how spatial properties of the dictionary change as a function of the degree of overcompleteness and sparsity imposed. Our main finding is that as either overcompleteness or sparsity is increased, new families of basis functions emerge, including multiscale

Further author information: (Send correspondence to BAO)
BAO: E-mail: baolshausen@berkeley.edu, Telephone: 510 642-7250
CFC: E-mail: cadieu@berkeley.edu
DKW: E-mail: dkwarland@ucdavis.edu, Telephone: 530 754-6670

Wavelets XIII, edited by Vivek K. Goyal, Manos Papadakis, Dimitri Van De Ville, Proc. of SPIE
Vol. 7446, 74460S · © 2009 SPIE · CCC code: 0277-786X/09/$18 · doi: 10.1117/12.825882

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 1/14/2018 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

blobs, ridge-like functions, and gratings. This finding is in line with recent work of Rehn & Sommer[7] who show that when "hard sparseness" is enforced on an overcomplete dictionary, diverse families of basis functions emerge that better reflect the actual diversity of receptive fields found in visual cortex.

Beyond sparsity, another important goal of representation is *stability.* That is, in representing a time-varying image we desire that the code variables change smoothly over time in a manner that directly reflects the structured transformations occurring in images. While the property of shiftability may be accommodated by an overcomplete representation, as mentioned above, it is by no means guaranteed. Thus, we must also consider how to adapt the dictionary and how to infer sparse representations in a manner that provides a smooth, continuous representation of image content amenable to interpretation at higher-levels of analysis. Here we show how this may be achieved by utilizing complex dictionaries composed of real and imaginary pairs. The resulting complex coefficients are factorized into amplitude and phase, which allows the amplitudes to be regularized in time to provide smooth, time-varying representations of image sequences.

As a result of the nonlinear factorization into amplitude and phase, new forms of structure are exposed in a simple form that may be exploited by additional layers of processing. In particular, the time-varying phase variables contain information about how features change over time, and the redundancies among these variables across the population may be exploited to extract higher-order dynamical properties from time-varying images. We demonstrate this fact by applying a second layer sparse coding model to the time-varying phase variables, showing that it is possible to learn transformations such as motion over a local region of the image. Thus, the complex, overcomplete encoding in the first layer provides a staging ground for extracting higher-order forms of structure in a second layer of processing.

In this paper we first describe our results learning sparse, overcomplete representations of static images using real-valued dictionaries. We then describe the extension to complex dictionaries and the factorization into amplitude and phase. Finally, we show how this factorization allows the learning of higher-order properties in a second stage of representation, and we discuss the implications of this scheme for hierarchical image coding.

2. LEARNING OVERCOMPLETE REPRESENTATIONS OF NATURAL IMAGES

As we have described previously,[5,6] one may adapt an overcomplete dictionary to the statistics of images by considering the elements or basis functions of the dictionary as parameters in a linear generative model of images:

$$I(\vec{x}) = \sum_{i=1}^{M} a_i \, \phi_i(\vec{x}) + \epsilon(\vec{x}) \tag{1}$$

where $I(\vec{x})$ denotes the set of pixel intensities in an image, $\vec{x}$ indexes position within the image, $\phi_i(\vec{x})$ are the basis functions of the dictionary and a_i their corresponding coefficients. The residual term $\epsilon(\vec{x})$ is included to capture structure not well described by the dictionary.

We desire a sparse representation in which only small fraction of a_i's need be non-zero to describe a given image, which we express mathematically by imposing a sparse, factorial prior over the coefficients,

$$P(\mathbf{a}) = \prod_{i=1}^{M} \frac{1}{Z_i} e^{-S(a_i)} \tag{2}$$

where the function S shapes the distribution over each coefficient to be peaked at zero with heavy tails. Here we use a Laplacian distribution, which gives $S(a_i) = \lambda_i \, |a_i|$, where λ_i is a constant scale factor. If we assume a Gaussian i.i.d. noise model on the residual $\epsilon(\vec{x}) \sim \mathcal{N}(0, \sigma_\epsilon^2)$, we obtain the following energy function to be minimized:

$$E = \frac{1}{2\sigma_\epsilon^2} \sum_{\vec{x}} \left[I(\vec{x}) - \sum_{i=1}^{M} a_i \, \phi_i(\vec{x}) \right]^2 + \sum_i \lambda_i \, |a_i| \tag{3}$$

In terms of the probabilistic model, E corresponds to the negative log-posterior over the coefficients, $P(\mathbf{a}|\mathbf{I};\Phi)$. Thus, minimizing E with respect to the a_i corresponds to computing the MAP estimate of the coefficients.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 1/14/2018 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Learning may also be accomplished by minimizing E with respect to the basis functions $\phi_i(\vec{x})$ averaged over many images. If we use the MAP-inferred coefficients in this latter minimization, then this procedure may be considered an approximation to maximizing the log-likelihood of the model. When the residual term is assumed to be zero and the number of dictionary elements M is less than or equal to the number of image pixels, this model is formally equivalent to so-called "ICA" (independent components analysis), as shown in [6].

Figure 1 shows the result of using this procedure to adapt a dictionary of 100 basis functions to a large set ($\sim 10^6$) of 16×16 pixel patches extracted from a database of 50 natural images.[8] The images are preprocessed by lowpass filtering and whitening in order to remove artifacts at the highest spatial-frequencies and to remove large anisotropies in variance due to the $1/f^2$ power spectrum of natural images (see [6]). The result of this preprocessing leaves approximately 100 independent dimensions within a 16×16 patch, and so we have chosen $M = 100$ here to give a critically sampled dictionary that we can use as a basis of comparison for the overcomplete dictionaries below. The parameters σ_ϵ and $\lambda_i \equiv \lambda \, \forall i$ were chosen to yield an overall SNR of 11 dB. The basis functions $\phi_i(\vec{x})$ were adapted via gradient descent on E over many images, using MAP-inferred coefficients computed via an iterative thresholding procedure.[9]

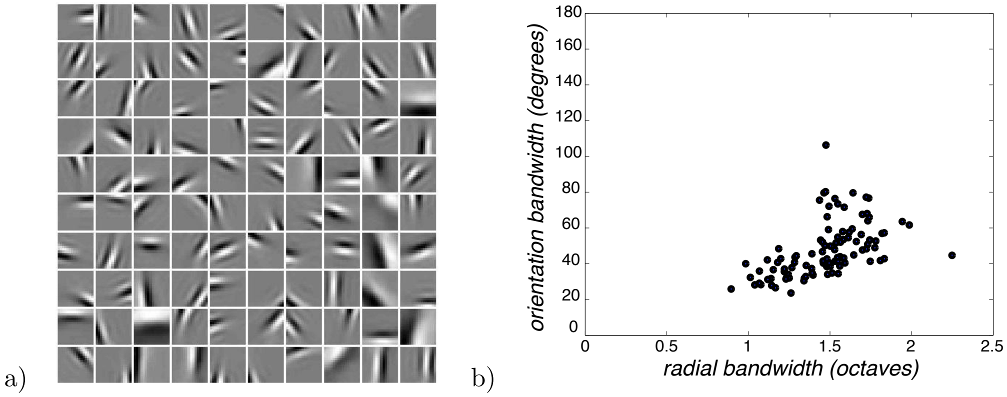

Figure 1. Learned basis functions (critically sampled). a) Full set of 100 basis functions learned on 16×16 pixel image patches. b) Distribution of basis functions as a function of spatial-frequency bandwidth and orientation bandwidth.

In line with previous results, the learned basis functions resemble Gabor wavelets in that they are localized, oriented, and multiscale with approximately equal bandwidth in octaves. Figure 1b shows the distribution of basis functions in spatial-frequency bandwidth and orientation bandwidth, which is clustered around 1.5 octave bandwidth and 40 deg. orientation bandwidth. This dictionary provides a sparsity fraction of 0.2 ($\langle |a_i|_{L0} \rangle = 20$) at this chosen SNR of 11 dB.

To explore how the dictionary changes with the degree of overcompleteness and sparsity, we learned dictionaries at three different levels of overcompleteness, 2.5x, 5x, and 10x, each with three different levels of sparsity imposed, 10%, 5%, and 1%. The level of sparsity was controlled by setting the parameters σ_ϵ and λ so as to yield the desired average number of non-zero coefficients (note that σ_ϵ and λ may be combined into a single parameter for this purpose). The results of this exploration are shown in Figure 2. One can see that as either the degree of overcompleteness or sparsity increases, the unimodal distribution around 1.5 octaves/40 deg. bandwidth breaks up into multiple classes. These include a set of narrowly orientation-tuned functions (lower portion of plot) and another that is non-orientation tuned (upper portion). These are best exemplified by the 10x dictionary at 1% sparsity (upper right plot of fig. 2), the full set of which is shown in Figure 3. Here one clearly sees three families of functions: 1) ridge or contour-like functions (lower right cluster in plot in fig. 2), 2) multiscale, non-oriented spots or blobs (upper cluster), and 3) gratings (lower left cluster). Some functions appear to describe curved contours (e.g., row 2, column 13), but could as easily be considered low spatial-frequency, bandpass blobs that are spatially truncated. It is clear in any case that expanding the degree of overcompleteness, or increasing the degree of sparsity, results in more diverse families of basis functions than what is seen in a critically sampled dictionary (fig. 1).

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 1/14/2018 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

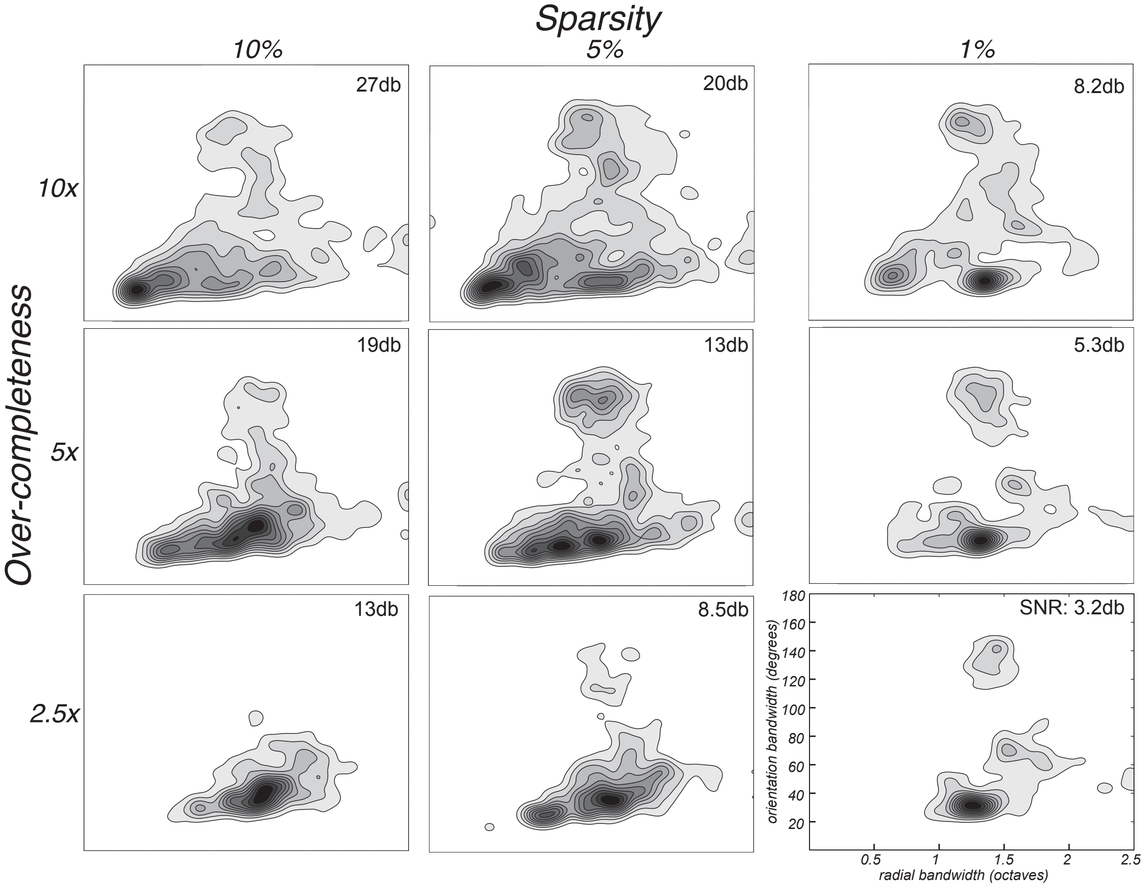

Figure 2. Distribution of basis functions as a function of spatial-frequency bandwidth (horizontal axis) and orientation bandwidth (vertical axis) for three different degrees of overcompleteness, each at three levels of sparsity. The level of sparsity is indicated by the percentage of non-zero coefficients. (A 1%, 10x representation has the same absolute number of non-zero coefficients as a 10%, 1x representation.) The resulting SNR for each combination of sparsity and overcompleteness is indicated in the upper right of each plot.

The emergence of new families of basis functions begs the question of whether each of these families forms a complete tiling within its own feature space. We do not have a rigorous answer to this question, but we attempt here to show how two families of basis functions, the non-oriented blobs and ridge-like functions, tile their respective domains. Figure 4a shows the position tiling of the non-oriented, blob-like functions at the highest spatial-frequency band, revealing that this class of functions provides a complete and fairly uniform coverage of the image patch. Figure 4b shows the ridge-like functions separated into six orientation bands, ordered by spatial position within each band. Again, the tiling appears fairly uniform across orientation and position with no obvious gaps in coverage. There appears to be a finer tiling of horizontal and vertical orientations which is expected due to the fact that the images used in training are photographs taken in the upright orientation. The grating-like functions are difficult to characterize in terms of their tiling as they have different spatial extents. From casual inspection they appear to provide only a sparse, incomplete tiling of the Fourier domain (spatial-frequency and orientation).

It seems plausible that the blob-like and ridge-like functions emerge to encode different forms of structure within the image. The latter would be well-suited for representing edges and contours, while the former would be suitable for representing 2D features such as corners, line terminations, or junctions. Note that if this were a linear code the blob-like functions would simply act as bandpass filters. However the sparse encoding involves

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 1/14/2018 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

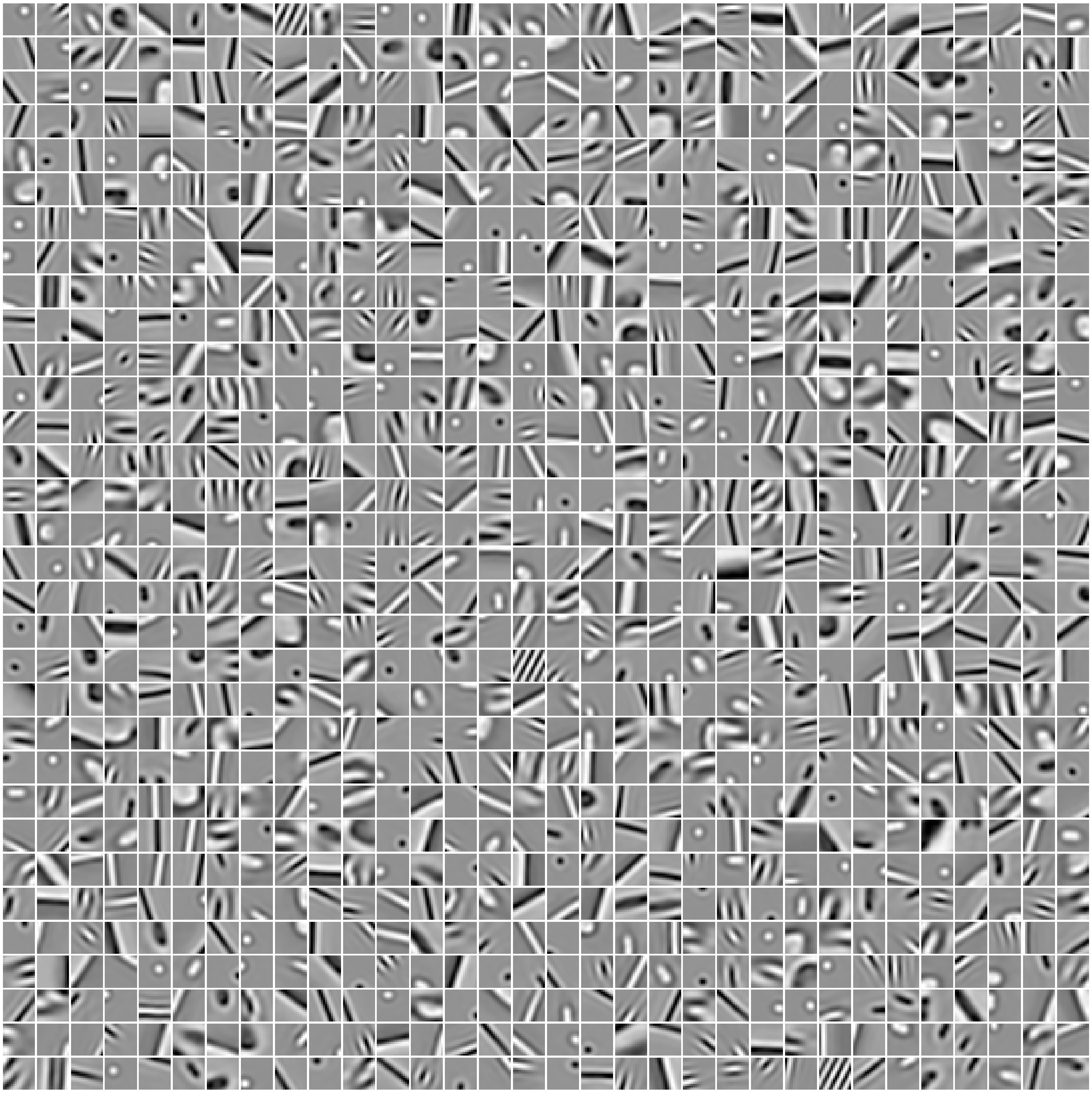

Figure 3. Full set of 1024 learned basis functions (10x overcomplete), corresponding to upper right corner of fig. 2. Three classes of functions are evident: 1) ridge or contour-like functions, 2) multiscale, non-oriented spots or blobs, and 3) gratings.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 1/14/2018 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

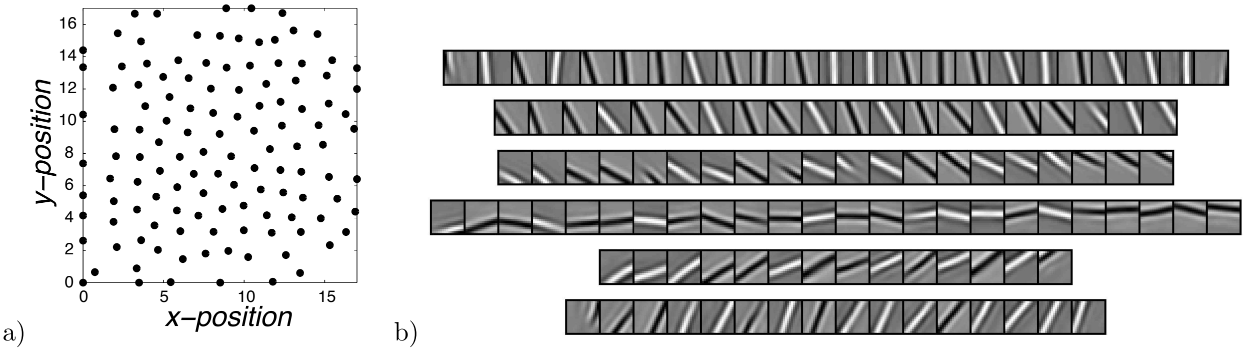

Figure 4. Tiling properties. a) Non-oriented, blob-like functions, highest spatial-frequency band only. Each dot indicates the center of mass within the image patch for each function. b) Ridge-like functions, separated into six orientation bands, sorted according to spatial position orthogonal to the axis of orientation within the patch.

a nonlinear selection process (minimization of eq. 3), so when the coefficient of a blob-like function is activated it indicates that this function is better suited in comparison to the overlapping ridge-like functions or gratings to encode the structure in this region of the image. Likewise, an elongated edge within the image will be better described by a ridge-like function, and so a blob-like function will not be activated in this case even though its inner product with the image may be substantial.

3. STABLE REPRESENTATION VIA AMPLITUDE AND PHASE FACTORIZATION

The learned dictionaries above enable a sparse representation of image content in terms of a set of spatial features matched to the statistics of the data. So far though we have said nothing about the *stability* of the representation. That is, how does the sparse encoding of an image change as the image changes? For the purpose of image analysis, video encoding, or object recognition, we desire that the coefficients change in a regular manner that reflects the underlying transformations in the image domain, but thus far there is nothing in the objective function of equation 3 that ensures or encourages this. Here we seek to regularize the temporal evolution of the coefficients by imposing a non-factorial prior over their activity that encourages groups of coefficients corresponding to the same feature to cooperate in a way that gives rise to smooth transitions in activity.

We may view the basis functions learned by the sparse coding model as a set of *interpolating functions*. That is, due to the linear generative model (eq. 1), any given feature within the image must be described by adding these functions together. If a feature matches one of the basis functions exactly, then only one coefficient need be activated to represent it fully. But more often than not an image feature will lie in the span between two or more basis functions. As the feature changes position, the activity of the corresponding coefficients must transition to code for these intermediate positions. This type of cooperative coding will give rise to a certain form of statistical dependency among coefficients—namely, a circularly symmetric, non-factorial distribution. Such distributions were first described by Zetzsche et al.,[10] who observed that the responses of pairs of Gabor functions in quadrature phase have a circularly symmetric, yet sparse (non-Gaussian), joint distribution. This structure suggests that pairs of such filter outputs are better described in polar coordinates—i.e., in terms of amplitude and phase—rather than cartesian coordinates (the responses of individual filers). In addition, Hyvarinen has pointed out that the way in which coefficients change over time in a sparse code exhibits a "bubble" like structure, which suggests factorizing the coefficients into a slowly changing amplitude envelope and a more rapidly changing variable.[11]

A natural way to factorize the coefficients into amplitude and phase is by utilizing a complex dictionary with complex coefficients. That is, we let the basis functions have real and imaginary parts, $\phi_i(\vec{x}) = \phi_i^R(\vec{x}) + j\phi_i^I(\vec{x})$, and the coefficients have amplitude and phase, $z_i = \sigma_i\, e^{j\,\alpha_i}$ $(j = \sqrt{-1})$. Multiplying $\phi_i(\vec{x})$ by z_i, and taking the real part of the product, allows us to interpolate between the real and imaginary parts according to the phase of z_i:

$$\Re\{z_i^*\,\phi_i(\vec{x})\} = \sigma_i\,[\cos\,\alpha_i\;\phi_i^R(\vec{x}) + \sin\,\alpha_i\;\phi_i^I(\vec{x})]$$

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 1/14/2018 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

where $\Re\{\}$ denotes 'real part.' Thus, we have constructed a 'phase shiftable' feature descriptor in which the amplitude σ_i indicates its presence or absence and is invariant with respect to some local transformation, and the phase α_i indicates the transformation. Note that this formulation is quite general in that a phase shift need not correspond to a literal shift in the image domain, but rather any transformation that may occur between features.

Now we can represent time-varying images using an overcomplete dictionary of such complex basis functions as follows:

$$\begin{aligned} I(\vec{x},t) &= \sum_{i=1}^{M} \Re\{z_i^*(t)\,\phi_i(\vec{x})\} + \epsilon(\vec{x},t) \\ &= \sum_i \sigma_i(t)\,[\cos\,\alpha_i(t)\;\phi_i^R(\vec{x}) + \sin\,\alpha_i(t)\;\phi_i^I(\vec{x})] + \epsilon(\vec{x},t) \end{aligned} \tag{4}$$

Note that one may equivalently view this as a collection of $2M$ basis functions with coefficients $u_i(t) = \sigma_i(t)\,\cos\alpha_i(t)$ and $v_i(t) = \sigma_i(t)\,\sin\alpha_i(t)$. Here though each pair of functions is modulated by a common amplitude and phase. This model may be adapted to time-varying natural image sequences by imposing both sparseness and *temporal smoothness* on the amplitudes, $\sigma_i(t)$, in order to encourage the model to learn real and imaginary pairs that correspond to the same feature related through a transformation in time. Thus we obtain the following energy function to be minimized for the new complex dictionary:

$$E = \sum_{\vec{x},t} \left[I(\vec{x},t) - \sum_i \sigma_i(t)\;\left[\cos\,\alpha_i(t)\;\phi_i^R(\vec{x}) + \sin\,\alpha_i(t)\;\phi_i^I(\vec{x})\right] \right]^2 + \sum_{i,t} \left[\lambda_s \sigma_i(t) + \lambda_t |\dot{\sigma}_i(t)|^2\right] \tag{5}$$

where the terms $\lambda_s\sigma_i(t)$ and $\lambda_t|\dot{\sigma}_i(t)|^2$ impose sparseness and temporal smoothness, respectively. Now the amplitudes and phases are computed by minimizing E with respect to $\sigma_i(t)$ and $\alpha_i(t)$. Since there is no penalty on the phase variables they will spin as needed in order to best match the time-varying structure in the image (sparsity in amplitude) and in a way that minimizes change in the amplitude. Learning is accomplished as before by minimizing E with respect to the complex dictionary $\{\phi_i\}$.

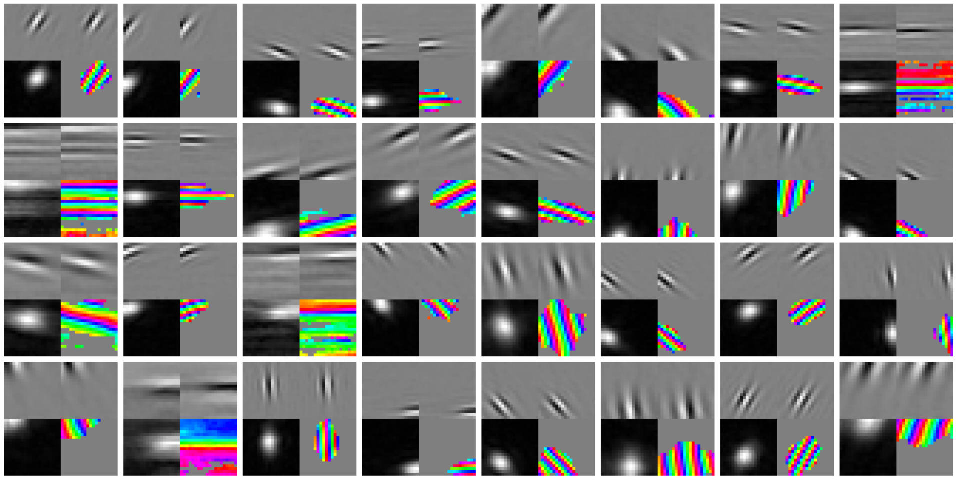

Figure 5. Learned complex basis functions. Shown is a random sampling of 32 complex basis functions (20 × 20 pixels) out of a dictionary of size 400. Each panel shows a real/imaginary pair (top) along with their complex modulus (lower left) and phase (lower right). The full dictionary, spinning in phase, may be viewed at `https://redwood.berkeley.edu/cadieu/pubs/videos/movie_TransInv_Figure2.mov`

Figure 5 shows a random sampling of complex basis functions learned as a result of adapting the model to natural image sequences (20 × 20 pixel image patches). These image sequences were extracted from nature

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 1/14/2018 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

documentaries obtained from Hans van Hateren's natural movie database.[8] The learned complex basis functions take on a similar form as before in the 1x representation (localized, oriented, bandpass), except now they come in Hilbert pairs—i.e., one function is the Hilbert transform of the other (this may be seen by taking the Fourier transform of each pair, revealing that they have one-sided spectra). This structure is not enforced, but rather emerges as a result of imposing sparsity and temporal smoothness on the amplitudes. When added together, weighted by the cosine and sine of the phase, α_i, each pair combines to form a shiftable basis function. One can see the range of variation expressed by each function by holding the amplitude of its coefficient fixed and spinning the phase from 0 to 2π (see complex basis function movie).

Figure 6 shows an encoding an image sequence using the complex basis function model. Note that the local invariances are now made explicit via the complex amplitudes, $\sigma_i(t)$, which are sparse but persist over time. By comparison, the real and imaginary coefficients, $u_i(t)$, $v_i(t)$, tend to undulate with each frame. In addition, motion is explicitly represented as a linear ramp in phase during the periods when the corresponding amplitude is significant.

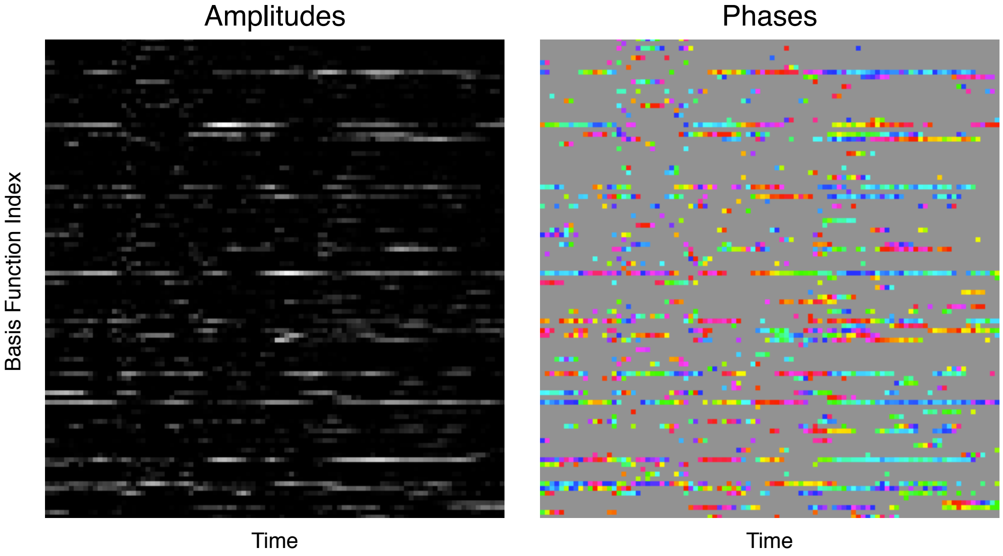

Figure 6. Coding of a 96-frame image sequence in terms of magnitude (left) and phase (right). Each row shows the activity of a different complex coefficient. Only a random sampling of 100 out of the total of 400 coefficients are shown. Phase is displayed on a continuous color scale only for those points in time when the corresponding amplitude is significant.

We may quantitatively compare the degree of temporal smoothness in the complex amplitudes $\sigma_i(t)$ as compared to the real and imaginary coefficients, $u_i(t)$, $v_i(t)$ via the power spectrum, as shown in figure 7a. The complex amplitudes are significantly lower pass in comparison to the real and imaginary coefficients. The phases by contrast precess over time as needed to encode how features change over time, as shown in figure 7b. Thus, we have obtained a sparse and temporally smooth representation of an image sequence in terms of locally shiftable features, with the amplitudes representing their contrast and the phases representing their spatial position.

4. LEARNING HIGHER-ORDER DYNAMICAL STRUCTURE

Given the decomposition into amplitude and phase variables, we now have a non-linear representation of image content that exposes higher-order structure in a simple form that may be exploited by a second layer of processing. In particular, the dynamics of objects moving in continuous trajectories through the world over short epochs will be encoded in the population activity of the phase variables α_i. Furthermore, because we have encoded these trajectories with an angular variable, many transformations in the image domain that would otherwise be

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 1/14/2018 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

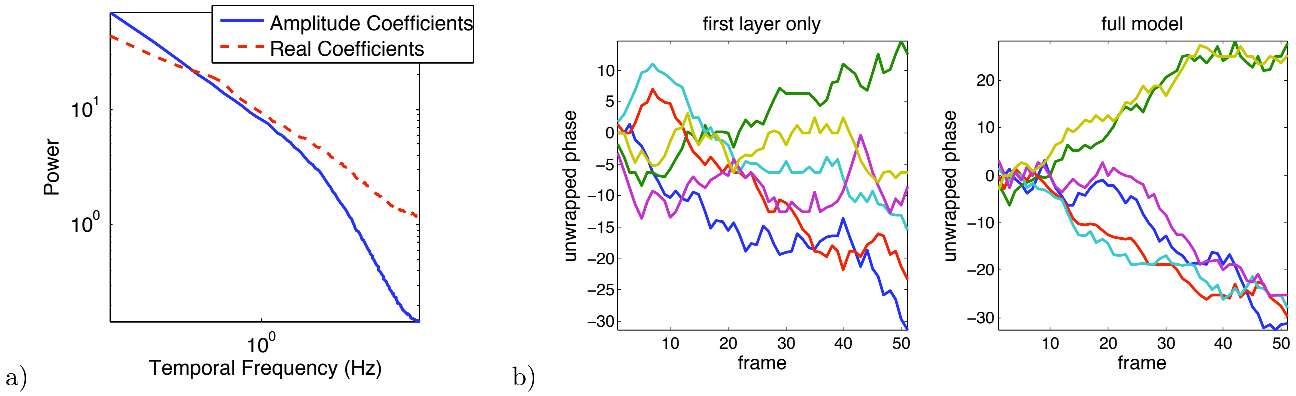

Figure 7. Amplitude and phase dynamics. a) Power spectrum of time-varying amplitudes in comparison to real and imaginary coefficients. b) Phase precession for six different complex coefficients in response to a 50 frame image sequence. 'first layer only' = non-informative prior over phase (eq. 5). 'full model' = second-layer prior over phase (eq. 8).

nonlinear in the coefficients u_i, v_i will now be linearized. This linear relationship allows us to model the time-rate of change of the phase variables, and hence the motion of objects, with a simple linear generative model.

We model the first-order time derivative of the phase variables $\dot{\alpha}_i$ as follows:

$$\dot{\alpha}_i(t) = \sum_k D_{ik}\, w_k(t) + \nu_i(t) \tag{6}$$

where $\dot{\alpha}_i = \alpha_i(t) - \alpha_i(t-1)$, and D is the basis function matrix specifying how the high-level variables w_k influence the phase shifts $\dot{\alpha}_i$. The additive noise term, ν_i, represents uncertainty or noise in the estimate of the phase time-rate of change. As before, we impose a sparse, independent distribution on the coefficients w_k, in this case with a sparse cost function given by:

$$S_w(w_{k(t)}) = \beta \log(1 + (\frac{w_k(t)}{\sigma})^2) \tag{7}$$

The uncertainty over the phase shifts is modeled as a von Mises distribution: $p(\nu_i) \propto \exp(\kappa \cos(\nu_i))$. Thus, the negative log-prior over the second layer units is given by

$$E_\alpha = \sum_t \sum_{i \in \{a_i(t) > 0\}} -\kappa \cos(\dot{\phi}_i - [Dw_{(t)}]_i) + \sum_k S_w(w_{k(t)}) \tag{8}$$

The structure of the complete graphical model is shown in Figure 8.

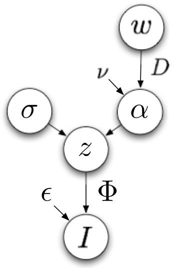

Figure 8. Graph of the hierarchical dynamical model showing the relationship among hidden variables.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 1/14/2018 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

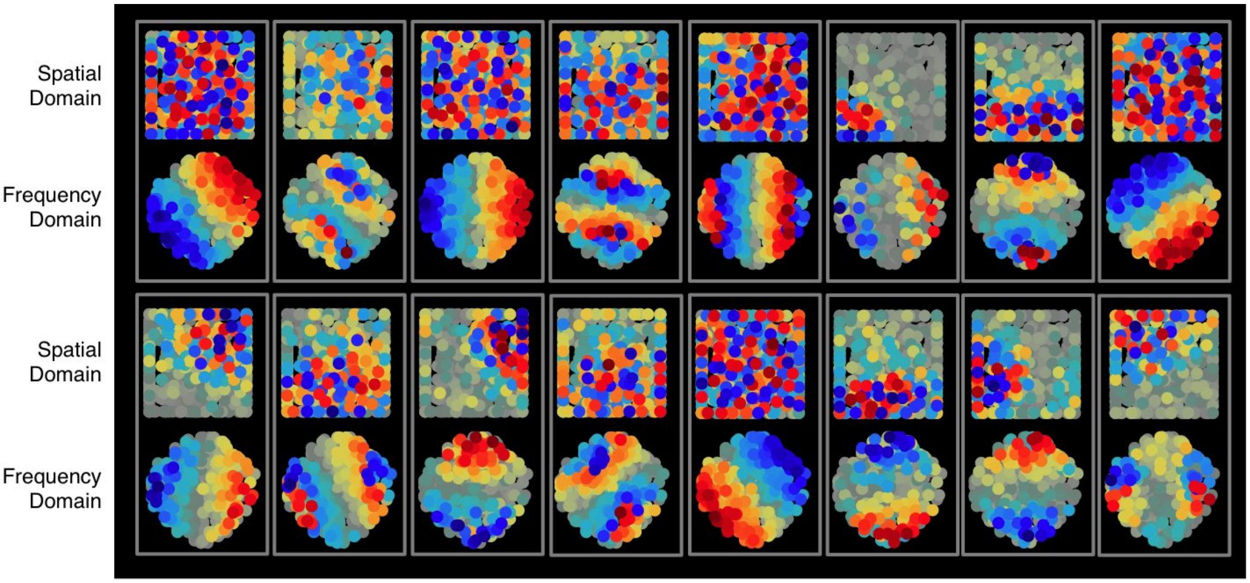

Figure 9. Learned second-layer weights D_{ij}. Shown are the weights of 16 second-layer units (out of a total of 100), each outlined by a white rectangle. For each, the top array shows the weights in the space domain and the bottom array shows the weights in the Fourier domain (with DC in the center). Each dot within an array corresponds to a complex unit in the first layer. The strength of each weight is denoted by hue (red +, blue -, gray 0).

Figure 9 shows a random sampling of 16 of the learned second layer weights, D_{ij}, visualized in both the space domain and frequency domain depictions of the first-layer units. Some have a global influence over all spatial positions within the 20x20 input array (e.g., row 1, column 1), while others have influence only over a local region (e.g., row 1, column 6). Those with a linear ramp in the Fourier domain (e.g., row 1, column 1) correspond to rigid translation, since the higher spatial-frequencies will spin their phases at proportionally higher rates (and negative spatial-frequencies will spin in the opposite direction). Some functions we believe arise from aliased temporal structure in the movies (row 1, column 5), and others are unknown (row 2, column 4). Details of the adaptation procedure for the weights D are given in a previous publication.[12]

The action of the second layer units may be best understood by activating one unit at a time and showing how it transforms the content of an image. For this purpose we encourage the reader to view these animations: movie 1, movie 2, movie 3, movie 4.

Another way to see the effect of the second layer is in the way it regularizes the phases over time, as shown in figure 7b. Without the second layer the phase precession over time can be rather erratic. With the second layer exerting an influence over the phase, via E_α (eq. 8), the phase precession becomes much more regular.

5. DISCUSSION

This paper provides two main contributions: 1) an exploration of how feature dictionaries change as a function of the degree of overcompleteness and sparsity, and 2) a method for regularizing the coefficients over time so as to provide sparse and smooth, time-varying representations of image sequences. In the first part we showed that new families of basis functions emerge as either the degree of overcompleteness or level of sparsity is increased. Two of these families—non-oriented blobs and oriented ridge-like functions—appear to completely tile their respective domains. Previous results by Rehn & Sommer[7] also showed that blob-like functions emerge as a result of imposing "hard sparseness," but they did not report such a prevalence of oriented ridge-like functions. However, they did not explore representations with the same degree of overcompleteness as we have here. Thus, it would appear that the additional degrees of freedom provided by a highly overcomplete dictionary allow the emergence of basis function families that are more specifically matched to features in natural images.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 1/14/2018 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

In the second part, we utilized a complex dictionary to allow factorization of coefficients into amplitude and phase. By imposing a temporal smoothness constraint on the amplitudes, we showed that it is possible to learn dictionaries of real and imaginary features related through a transformation. However, no blob or ridge-like functions emerge from these complex dictionaries. Even with complex dictionaries up to 4x overcomplete (results not shown), the learned features remain Gabor-like, without the diverse families seen with real dictionaries on static images. Thus, it may be that the diverse families of features learned with real dictonaries do not generalize well to provide smooth representations of image sequences.

The representation of an image sequence in terms of time-varying amplitude and phase variables makes it possible for neurons in a second layer to extract higher-order properties using a simple, linear generative model. In this context, the overcomplete representation provided by the first layer provides a staging ground for extracting higher-order structure in additional layers. Conceivably, this architecture could be recapitulated in multiple layers to provide a hierarchical description of images.

Thus far our strategy has been mainly focused on pure, unsupervised learning, which does not adjudicate between different types of information but merely tries to reformat data in a way the makes explicit the forms of structure contained within it. Such representations may be useful for myriad tasks, but ultimately, specific task demands will dictate what forms of structure need to be extracted and which can be discarded. Thus, an important goal of future work will be to combine bottom up unsupervised learning with top-down task constraints to develop an optimal low level coding scheme.

ACKNOWLEDGMENTS

The authors wish to thank Fritz Sommer and Martin Rehn for early collaborative efforts[13] and helpful discussions characterizing the diversity of the learned basis functions. Work supported by grants from the National Geospatial Intellgence Agency (HM1582-08-1-0007), the National Science Foundation (IIS-0705939) and the Canadian Institute for Advanced Research.

REFERENCES

[1] Simoncelli, E., Freeman, W., Adelson, E., and Heeger, D., "Shiftable multiscale transforms," *Information Theory, IEEE Transactions on* **38**(2), 587–607 (1992).

[2] Doi, E., Balcan, D. C., and Lewicki, M. S., "Robust coding over noisy overcomplete channels," *IEEE Transactions on Image Processing* **16**, 442–52 (Feb 2007).

[3] Schmid-Saugeon, P. and Zakhor, A., "Dictionary design for matching pursuit and appliation to motion compensated video coding," *IEEE Transactions on Circuits and Systems for Video Technology* (6), 880 – 886 (2004).

[4] Rahmoune, A., Vandergheynst, P., and Frossard, P., "Flexible motion-adaptive video coding with redundant expansions," *IEEE Transactions on Circuits and Systems for Video Technology* **16**(2), 178–190 (2006).

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

[6] Olshausen, B. A. and Field, D. J., "Sparse coding with an overcomplete basis set: a strategy employed by V1?," *Vision Research* **37**(23), 3311–25 (1997).

[7] Rehn, M. and Sommer, F. T., "A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields," *J Comput Neurosci* (Oct 2006).

[8] van Hateren, H., "Natural stimuli collection," *http://hlab.phys.rug.nl/archive.html* .

[9] Rozell, C. J., Johnson, D. H., Baraniuk, R. G., and Olshausen, B. A., "Sparse coding via thresholding and local competition in neural circuits," *Neural Computation* **20**, 2526–2563 (2008).

[10] Zetzsche, C., Krieger, G., and Wegmann, B., "The atoms of vision: Cartesian or polar?," *Journal of the Optical Society of America A* **16**(7), 1554–1565 (1999).

[11] Hyvarinen, A., Hurri, J., and Vayrynen, J., "Bubbles: A unifying framework for low-level statistical properties of natural image sequences," *Journal of the Optical Society of America* **20**(7), 1237–1252 (2003).

[12] Cadieu, C. F. and Olshausen, B. A., "Learning transformational invariants from natural movies," *Advances in Neural Information Processing Systems, 21* , 209–216 (2009).

[13] Rehn, M., Warland, D. K., and Sommer, F. T. *Society for Neuroscience Abstracts* , 394.6 (2007).

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 1/14/2018 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use