ARTICLE Communicated by Aapo Hyvärinen

Learning Intermediate-Level Representations of Form and Motion from Natural Movies

Charles F. Cadieu
cadieu@berkeley.edu
Bruno A. Olshausen
baolshausen@berkeley.edu
Redwood Center for Theoretical Neuroscience, Helen Wills Neuroscience Institute, and School of Optometry, University of California, Berkeley, Berkeley, CA 94720, U.S.A.

We present a model of intermediate-level visual representation that is based on learning invariances from movies of the natural environment. The model is composed of two stages of processing: an early feature representation layer and a second layer in which invariances are explicitly represented. Invariances are learned as the result of factoring apart the temporally stable and dynamic components embedded in the early feature representation. The structure contained in these components is made explicit in the activities of second-layer units that capture invariances in both form and motion. When trained on natural movies, the first layer produces a factorization, or separation, of image content into a temporally persistent part representing local edge structure and a dynamic part representing local motion structure, consistent with known response properties in early visual cortex (area V1). This factorization linearizes statistical dependencies among the first-layer units, making them learnable by the second layer. The second-layer units are split into two populations according to the factorization in the first layer. The form-selective units receive their input from the temporally persistent part (local edge structure) and after training result in a diverse set of higher-order shape features consisting of extended contours, multiscale edges, textures, and texture boundaries. The motion-selective units receive their input from the dynamic part (local motion structure) and after training result in a representation of image translation over different spatial scales and directions, in addition to more complex deformations. These representations provide a rich description of dynamic natural images and testable hypotheses regarding intermediate-level representation in visual cortex.

1 Introduction

A key attribute of visual perception is the ability to extract invariances from visual input. How this is accomplished by neural circuits in the visual cortex

 © 2011 Massachusetts Institute of Technology

has been the subject of intense investigation in neuroscience over the past several decades. From this body of work, we know that visual information is processed incrementally in a series of cortical stages: neurons at early levels such as V1 appear to represent local features such as orientation and motion (Hubel & Wiesel, 1968), while neurons at higher levels such as in the inferotemporal and posterior parietal regions represent more global properties such as object identity (Kobatake & Tanaka, 1994; Tsunoda, Yamane, Nishizaki, & Tanifuji, 2001; Yamane, Carlson, Bowman, Wang, & Connor, 2008) and complex motion trajectories (Born & Bradley, 2005; Orban, 2008). These findings have led to a number of theoretical formulations of the problem (Gibson, 1983; Olshausen, Anderson, & Van Essen, 1993; Riesenhuber & Poggio, 1999; Ullman, 2000; DiCarlo & Cox, 2007). However, the precise nature of neural computation that takes place in intermediate-level areas to enable this transformation remains a mystery. Our theoretical approach to intermediate-level vision incorporates factorization of the visual world's constituent causes and learning the natural statistics of these causes.[1] Under this theoretical approach, we implement a model for how neurons in intermediate-level areas could factor movies into invariances related to form and motion, and we adapt the representation of these invariances to the statistics of the natural environment.

Central to our approach is the hypothesis that biological sensory systems are adapted to the statistics of their input (Barlow, 1961; Field, 1994). In previous modeling work, Olshausen and Field (1996) showed that when a neural system is adapted to the statistics of natural images so as to produce a sparse representation, the receptive fields that emerge are localized, oriented, and multiscale, in line with the response properties of simple cells in primary visual cortex. Such a representation is advantageous because it makes explicit the local features occurring in natural images. However, the underlying causes of form and motion are still entangled: as an object moves over the input array, the activity of neurons will be sparse but will also fluctuate dramatically as features of the object move in and out of the receptive fields of individual neurons. What is desired is a representation in which form and motion are disentangled. That is, some units should represent the form of the object in a persistent manner that is independent of motion, while other units represent the dynamics or motion that the object is undergoing independent of its form. Thus, neurons are selective for either form or motion information, but they are also invariant to the complementary aspect of visual information (motion or form, respectively). We use the terms *form-selective invariances* and *motion-selective invariances* to refer to these complementary aspects of visual representation.

[1] While we believe that a complete theory of intermediate-level vision will include representations of surfaces and grouping processes, we believe that invariance, learning, and factorization are also major aspects of intermediate-level vision.

There have been numerous efforts to learn form-selective invariances from the statistics of natural images (Einhauser, Kayser, Konig, & Kording, 2002; Karklin & Lewicki, 2005, 2008; Hyvärinen, Hurri, & Väyrynen, 2003; Lee, Ekanadham, & Ng, 2007), especially with the goal of producing representations that are useful for object recognition (Wallis & Rolls, 1997; LeCun, Huang, & Bottou, 2004; Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007). Many statistical approaches to this problem capture dependencies among oriented filter responses and reproduce properties of complex cells in primary visual cortex (Einhauser et al., 2002; Hyvärinen et al., 2003; Koster & Hyvärinen, 2007; Karklin & Lewicki, 2008; Berkes, Turner, & Sahani, 2009) or higher areas (Hoyer & Hyvärinen, 2002). Our approach leverages some aspects of this body of work, especially the idea of slow feature analysis as a principle of self-organization (Wiskott & Sejnowski, 2002). Models that learn form-selective invariances and focus on performance evaluation of object recognition tasks (Wallis & Rolls, 1997; LeCun et al., 2004; Serre et al., 2007) often have the specific invariance built in to the model structure, and the higher-order features that emerge beyond these built-in invariances have not been explored. The model we propose here bears similarities to the density components model of Karklin and Lewicki (2005) and to the hierarchical GSM model of Schwartz, Sejnowski, and Dayan (2006), which learn higher-order structure in images by modeling the dependencies in scale among oriented filter responses. Our model differs in that form-selective invariances are learned from the temporally persistent structure contained in natural movies as opposed to static image patches. In addition, our model captures dependencies among the normalized filter responses that remain after the scale has been divided out, which we parameterize as *phase*. These differences allow our model to learn forms of higher-order structure beyond those previously reported.

In contrast to learning of form-selective invariants, there has been little work on learning motion-selective invariances from natural image sequences. Previous efforts have relied on using either unnatural motions generated by hand (Nowlan & Sejnowski, 1995; Zhang, Sereno, & Sereno, 1993; Rolls & Stringer, 2007) or unrealistic supervised learning algorithms accounting for only rigid global translation of an image (Grimes & Rao, 2005). In other models (Jhuang, Serre, Wolf, & Poggio, 2007), it is not clear if they have captured the diversity of naturally occurring movements or if the representations are invariant to visual form. Another closely related line of work learns sparse, spatiotemporal representations of image sequences (Olshausen, 2002). This model produces local, direction-selective components that capture key aspects of measured space-time receptive fields in primary visual cortex. However, this type of model does not capture the abstract, invariant property of motion because each unit is bound to a specific orientation, spatial frequency, and location within the image. Like the sparse codes learned on static images, motion and form are still entangled.

The key attribute of our model that sets it apart from these previous approaches is that it jointly estimates form and motion by factoring the time-varying pixel data into persist and dynamic components. This approach stands in contrast to traditional models of form and motion processing in which these properties are extracted using computations that take place in separate, independent streams in visual cortex (Simoncelli & Heeger, 1998; Serre et al., 2007). Such approaches do not exploit the fact that information about form and motion is bound together in the incoming sensory data and that recovery of one of these properties depends on knowing the other: Computing the true motion relies on knowing the spatial pattern being compared across time, but since the pattern is not initially known, it must also be estimated from the time-varying image by integrating evidence over time in the motion-transformed pattern. Thus, a better estimate of one factor improves the estimate of the other. This principle was recently exploited in a model of retinal image motion compensation (Burak, Rokni, Meister, & Sompolinsky, 2010).

In this article, we present a hierarchical, probabilistic generative model for learning form- and motion-selective invariances from the statistics of natural movies. We begin by describing the overall structure of the model and the role of factorization. We then describe the first layer of the model, which uses a sparse coding model composed of complex basis functions, and we show how it provides a factorization into amplitude and phase that linearizes statistical dependencies. Next, we describe the second layer of the model, which learns from the factorized representation in the first layer, and we describe the form- and motion-selective invariances that emerge when trained on natural movies. In section 5, we relate our work to other models of visual form and motion processing, discuss the implications of factorization models for form and motion processing in visual cortex, and describe the limitations of our approach.

2 Model Overview

The model consists of an input layer and two hidden layers, as shown in Figure 1. The input to the model consists of the time-varying image pixel intensities $I(t)$. Local features of the image data are represented in the first layer, and these are grouped into representations of form and motion in the second layer. The weights between layers A, B, and D are learned by adapting to the statistics of natural movies.

The first hidden layer is a sparse coding model utilizing complex basis functions A and shares properties with independent subspace analysis (Hyvärinen & Hoyer, 2000) and the standard energy model of complex cells (Adelson & Bergen, 1985). The corresponding complex coefficients are represented in terms of amplitude $a(t)$ and complex phasor $e^{j\phi(t)}$. Sparseness and temporal persistence are imposed on the amplitudes $a(t)$. The basis functions A are then adapted to the statistics of natural movies so as to best

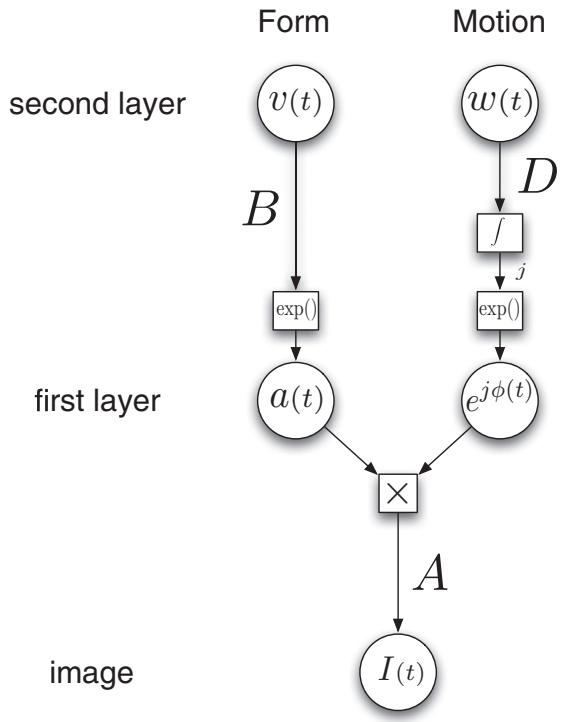

Figure 1: Model architecture. The first hidden layer is a sparse coding model utilizing complex basis functions A. The corresponding complex coefficients are factorized into amplitude $a(t)$ and a complex phasor $e^{j\phi(t)}$. The second hidden layer is a sparse coding model of the logarithm of the quantities represented in the first layer. In the motion pathway, the second layer models the time derivative of phase. The higher-order form and motion components, B and D, are learned from the statistical dependencies contained in the amplitudes and phase derivatives inferred from natural movies. These learned form and motion components are represented by the second-layer variables $v(t)$ and $w(t)$, respectively.

represent images under this constraint. Sparsity forces the basis functions to align with features contained in images, while temporal persistence encourages each complex pair to span a local manifold of the data such that image changes are better described in terms of phase shifts as opposed to changes in amplitude. In this way, the amplitudes capture local invariances in the image (form), while the phases capture local transformations (motion).

The second hidden layer is a sparse coding model of the logarithm of the quantities represented in the first layer. As we shall see, the logarithm,

in combination with the factorization into amplitude and phase, linearizes statistical dependencies among the first-layer variables, thus making them learnable by another linear generative model. The second-layer weights B thus capture the higher-order form structure contained in the log amplitudes, while the secondlayer weights D capture the higher-order motion structure contained in the time derivative of first-layer phases. These higher-order form and motion components, which we refer to synonymously as *amplitude components* and *phase-shift components*, are represented by the second-layer variables $v(t)$ and $w(t)$, respectively.

Because the hierarchical generative model we propose multiplies signals related to form and motion, the inference of these properties depends on each other and can be considered as a factorization problem (Tomasi & Kanade, 1992; Koenderink & Van Doorn, 1997; Memisevic & Hinton, 2007). Here, the inference of amplitudes in the form pathway depends on the transformation enabled by phase shifting in the motion pathway. At the same time, the inference of phase in the motion pathway is guided by the desire to achieve a sparse and stable representation of amplitude in the form pathway. Thus, form and motion are not computed independently. Our first-layer factorization approach is similar to that of Berkes et al. (2009), in which identity and appearance are factored; however, here we leverage this factorization with a hierarchical model that learns higher-level properties of form and motion, which are relevant to intermediate-level visual representation.

3 First Layer: Sparse and Temporally Persistent Representation

Previous work has shown that many of the observed response properties of neurons in V1 may be accounted for in terms of a sparse coding model of images (Olshausen & Field, 1997; Bell & Sejnowski, 1997):

$$I_{(x,t)} = \sum_i u_{i(t)} A_{i(x)} + n_{(x,t)}, \tag{3.1}$$

where $I_{(x,t)}$ is the image intensity as a function of space ($x \in \mathcal{R}^2$) and time ($t \in \mathcal{R}_+$), $A_{i(x)}$ is a spatial basis function with coefficient u_i, and the term $n_{(x,t)}$ corresponds to gaussian noise with variance σ_n^2 that is small compared to the image variance. Sparsity is imposed by a kurtotic, independent prior over the coefficients, $P(u) = \Pi_i \frac{1}{Z_i} e^{-S(u_i)}$, where S is typically chosen to correspond to either a Laplacian or Cauchy distribution. When adapted to an ensemble of image patches extracted from natural scenes, the $A_{i(x)}$ converge to a set of localized, oriented, multiscale functions similar to a Gabor wavelet decomposition of images.

Here we generalize the sparse coding model to complex variables (Cadieu & Olshausen, 2009). This step is motivated by a number of findings from natural scene statistics, neurophysiology, and human psychophysics

and draws heavily from the work of Christoph Zetzsche described in Zetzsche, Krieger, & Wegmann (1999). In particular, we are motivated by two observations: The first is that although the prior over the coefficients in sparse coding models is typically factorial, the actual joint distribution of coefficients, even after learning, exhibits strong statistical dependencies in response to natural images. One particularly prevalent form of dependency is a circularly symmetric, yet kurtotic, distribution found among pairs of coefficients with basis functions at nearby spatial positions, scales, or orientations (Wegmann & Zetzsche, 1990). Such a circularly symmetric distribution strongly suggests that these pairs of coefficients are better described in polar coordinates rather than Cartesian coordinates—that is, in terms of amplitude and phase. The second observation comes from considering the dynamics of coefficients through time. As Hyvärinen et al. (2003) pointed out, the temporal evolution of a coefficient, $u_{i(t)}$, in response to a movie can be well described in terms of the product of two variables: a smooth or temporally persistent amplitude envelope multiplied by a quickly changing variable akin to a carrier. A similar result from Einhauser et al. (2002) indicates that temporal continuity in amplitude provides a strong cue for learning local invariances. These results are closely related to the trace learning rule of Foldiak (1991) and slow feature analysis (Wiskott & Sejnowski, 2002; Berkes & Wiskott, 2005), which attempt to extract slowly changing signals from time-varying input.

In addition to these theoretical considerations, it should be noted that simple cells in V1 are highly selective to the local phase of an oriented edge and show contrast invariance above a saturation value (Albrecht & Geisler, 1991). Such responses are not in agreement with purely linear models. These nonlinearities suggest that primary visual cortex employs a coding strategy that is polar separable and not Cartesian separable (as would be assumed in linear models of simple cell responses). Furthermore, phenomenological models of simple cells, such as divisive normalization (Heeger, 1991), also result in similar response profiles and tuning that is selective for phase and invariant to contrast or amplitude. Other statistical models of natural images that are linked to neural responses also impose divisive normalization of linear responses (Schwartz & Simoncelli, 2001). These observations at the neurobiological level are supported by human psychophysics. The sensitivity of human observers to quadrature pair Gabor stimuli is aligned with a polar decomposition and not a Cartesian decomposition (Zetzsche et al., 1999). In sum, these observations strongly advocate the use of angular decompositions, such as those that can be obtained with complex variables, in models of natural images.

With these observations in mind, we have modified the sparse coding model by using a complex basis function decomposition as follows:

$$I_{(x,t)} = \sum_i \Re\{z_i^*(t)\, A_{i(x)}\} + n_{(x,t)}, \tag{3.2}$$

where the basis functions now have real and imaginary parts, $A_{i(x)} = A_i^{\mathcal{R}}{}_{(x)} + \mathbf{j}A_i^{\mathcal{I}}{}_{(x)}$, and the coefficients are also complex, with $z_{i(t)} = a_{i(t)}e^{\mathbf{j}\phi_{i(t)}}$. ($*$ indicates the complex conjugate, and the notation $\Re\{.\}$ denotes taking the real part of the argument.) Note that the resulting generative model can also be written as

$$I_{(x,t)} = \sum_i a_{i(t)} \left[\cos \phi_{i(t)} A_i^{\mathcal{R}}{}_{(x)} + \sin \phi_{i(t)} A_i^{\mathcal{I}}{}_{(x)}\right] + n_{(x,t).} \tag{3.3}$$

Thus, each pair of basis functions, $A_i^{\mathcal{R}}$ and $A_i^{\mathcal{I}}$, forms a two-dimensional subspace and is controlled by a common amplitude a_i and phase ϕ_i that determine the radius and angle within each subspace. Note that the basis functions are only functions of space. Therefore, the temporal dynamics within image sequences will be expressed in the temporal dynamics of the amplitude and phase. The relationship to the original sparse coding model, equation 3.1, can be seen by defining variables $u_i^{\mathcal{R}} = a_i \cos \phi_i$ and $u_i^{\mathcal{I}} = a_i \sin \phi_i$, which are the coefficients of the basis functions $A_i^{\mathcal{R}}$ and $A_i^{\mathcal{I}}$, respectively.

The prior over the complex coefficients, z, is designed so as to favor circularly symmetric distributions and smooth-amplitude dynamics as observed in time-varying natural images. As observed empirically (Hyvärinen et al., 2003), we expect the structural image content to be persistent through time or slowly changing. The prior we choose to enforce these constraints is

$$P(a_{i(t)}|a_{i(t-1)}) \propto e^{-\lambda_a Sp_a(a_{i(t)}) - \beta_a Sl_a(a_{i(t)}, a_{i(t-1)})}. \tag{3.4}$$

The first term in the exponential imposes a sparse prior on the coefficient amplitudes. Here we use a Cauchy prior on the amplitude variables:

$$Sp_a(a_{i(t)}) = \log\left(1 + \left(\frac{a_{i(t)}}{\sigma}\right)^2\right). \tag{3.5}$$

Other kurtotic priors yield similar results. Because the prior over the phases is uniform, the prior for each subspace specifies a circularly symmetric kurtotic distribution. The second term in the exponential imposes temporal stability on the time rate of change of the amplitudes and is given by

$$Sl_a(a_{i(t)}, a_{i(t-1)}) = (a_{i(t)} - a_{i(t-1)})^2. \tag{3.6}$$

Here we have chosen the l_2 distance, but we have found similar results for the l_1 distance, which would allow for sharp changes in motion. This prior also assumes that the temporal dependencies on the amplitudes form a Markov chain in time. The multiplicative factors λ_a and β_a control the relative influence of the sparsity and temporal continuity terms.

For a sequence of images, the resulting negative log posterior, or energy function, for the first hidden layer becomes (excluding an additive constant)

$$
\begin{aligned}
-\log P(I, a, \phi) \propto \quad & \sum_{x,t} \tfrac{1}{\sigma_N^2} \left[I_{(x,t)} - \sum_i \Re\{z_i^*{}_{(t)} A_{i(x)}\} \right]^2 + \\
& \lambda_a \sum_{i,t} Sp(a_{i(t)}) + \\
& \beta_a \sum_{i,t} Sl(a_{i(t)}, a_{i(t-1)}).
\end{aligned} \tag{3.7}
$$

The amplitudes a and phases ϕ of the first hidden layer are computed by minimizing this function through a gradient descent procedure (see the appendix). Note that since there is no prior over the phases, they will essentially steer each basis function over time so as to achieve a sparse and temporally persistent representation in the amplitudes. The basis functions are adapted to the statistics of image sequences by following the gradient of this same energy function using the inferred amplitudes and phases, as described in the appendix. Thus, we are essentially asking the system to learn a set of trackable features matched to the structure of images.

While this model by no means captures the full joint distribution of coefficients, it does at least capture the circular symmetric dependencies among local groups (pairs) of coefficients, which allows for the explicit representation of amplitude and phase. As we shall see, this nonlinearity in the form of a multiplicative interaction between the amplitude and phasor variables serves as a staging ground for learning higher-order dependencies over space and time.

After training on an ensemble of natural movies (see the appendix), the first-layer complex basis functions converge to a set of localized, oriented, and bandpass functions. Figure 2a shows 16 randomly selected 32×32 pixel basis functions in terms of both their real and imaginary parts and in terms of their amplitude and phase. Each pair exhibits similar tuning in position, orientation, and spatial frequency, in line with previous results where temporal persistence is enforced on group amplitudes (Einhauser et al., 2002; Hyvärinen et al., 2003; Berkes et al., 2009). We can also examine the joint amplitude and phase of each complex pair. Here we see that the amplitude envelopes are well localized and have a roughly gaussian profile, while the phase reveals a smooth ramp in the direction perpendicular to the basis functions' orientation. Note that each pair has converged to a roughly quadrature-pair phase relationship in which the spatial phase of each element of a pair is shifted by 90 degrees. This relationship was not enforced; it emerges from the data.

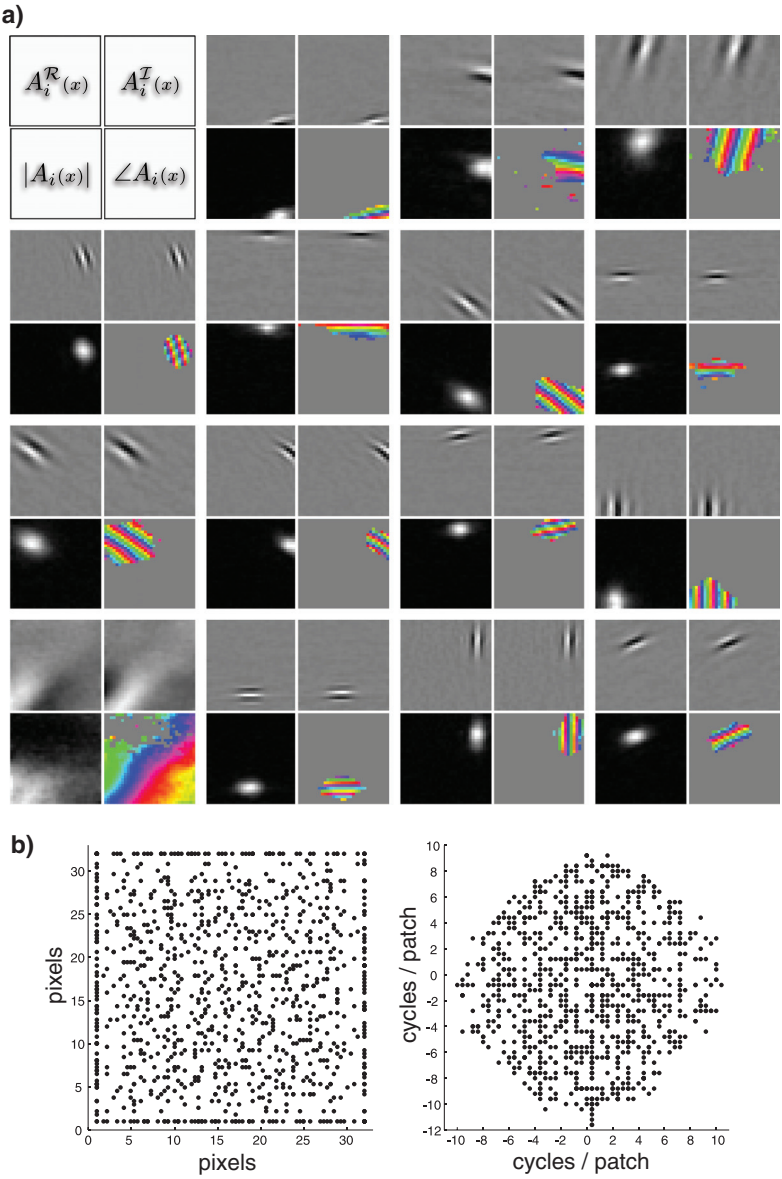

Figure 2b shows how the entire population of 1024 complex basis functions tiles spatial position (left) and spatial frequency (right). Each dot represents a different basis function according to the location of its maximum amplitude in the space domain, or the location of its maximum amplitude

in the frequency domain computed via the 2D Fourier transform of each complex pair. Because each basis function is complex and in quadrature, it produces only a single peak in the spatial frequency plane. The angle of each dot in the spatial frequency plane is determined by the orientation of the complex basis function pair, and the radius is determined by the magnitude of the dominant spatial frequency. Basis functions with low spatial frequency are represented near the origin, and those with high spatial frequencies are located far from the origin. Similar to previous models (Olshausen & Field, 1997; Van Hateren & Van der Schaaf, 1998; Karklin & Lewicki, 2006), the basis functions uniformly tile both domains. Note that because we consider the basis function's position to be determined by the position of the maximum, there is a bias in the space domain toward the boundaries (all basis functions that have their true maximum of the envelope outside the image patch are projected to the edge of the image patch). The coverage of the spatial-frequency plane spans spatial frequencies up to the rolloff of the lowpass filter used in preprocessing. Although we trained the first layer and second layer separately, we observe a uniform and dense tiling of the spatial frequency plane. This finding does not match that of Karklin & Lewicki (2006), which found it necessary to learn a second-layer model jointly with the first-layer model to achieve such uniform tiling. This difference may be due to the subspaces we introduce in the complex basis functions or to a difference in the data sets or learning procedures.

The factorization of the complex basis function coefficients into amplitude and phase provides a staging ground for separating form and motion structure. We demonstrate the effect of factorization in Figure 3 by showing

Figure 2: Learned first-layer basis functions. (a) Each 2 × 2 panel shows a learned complex function over the space domain in terms of both its real and imaginary parts, $A_i^{\mathbf{R}}(x)$, $A_i^{\mathbf{I}}(x)$, and complex modulus and phase, $|A_i(x)|$, $\angle A_i(x)$, as shown in the legend at the upper left. Real and imaginary values are displayed on a gray-scale color map with zero mapped to the middle of the scale and the range normalized independently for each complex basis function to span the maximum range of gray values without clipping. Complex modulus is displayed on a gray-scale map where a value of 0 corresponds to black and a normalized value of 1 corresponds to white. Complex phase is displayed using a circular color map so as not to produce wrap-around discontinuities. Phase is displayed as gray where the corresponding amplitude falls below 10% of its maximum value (where phase is not well defined). See the corresponding animation available online at http : //www.vimeo.com/album/1624584. (b) Tiling of space and spatial frequency for the entire population of basis functions. Each dot in either the space domain (left) or spatial-frequency domain (right) corresponds to a learned basis function. Grid clustering is due to quantization of the Fourier domain. This representation of the first-layer basis functions is used to visualize the weights learned in the second layer.

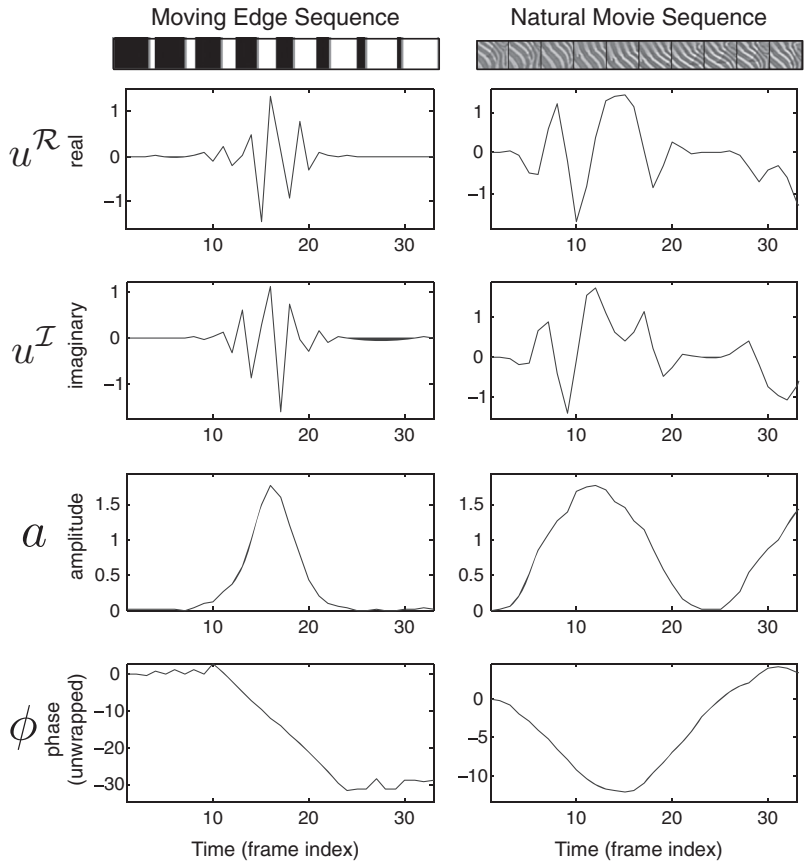

Figure 3: Factorization into amplitude and phase. The first-layer model factors visual content into amplitude a and phase ϕ variables, which are more directly related to image form and motion as compared to the linear coefficients $u^{\mathcal{R}}$ and $u^{\mathcal{I}}$. The left column shows the evolution of these variables for a sharp edge moving horizontally across the visual field (subsampled sequence shown above first plot). The right column shows the evolution of these variables for a natural movie sequence containing complex motion.

the inferred latent variables for a sharp edge moving horizontally across the image patch (left column) and for a natural movie sequence containing complex motion (right column). We compare the evolution of linear coefficients $u^{\mathcal{R}}$ and $u^{\mathcal{I}}$ to the amplitude a and phase ϕ (unwrapped through time) for one complex basis function. The linear coefficients $u^{\mathcal{R}}$ and $u^{\mathcal{I}}$ exhibit complex trajectories through time, while the amplitude and phase follow smooth trajectories. The magnitude of the amplitude indicates the presence

of the feature within the image, while the phase is related to the absolute position of the edge within the image. Importantly, the time derivative of phase is directly related to the speed at which the edge moves through space.[2] Thus, the amplitudes and phases effectively separate, or factor, the presence of edge structure from the movement of the edge structure.

4 Second Layer: Representation of Form and Motion

The goal of the second layer is to provide an efficient representation of structure contained in the amplitudes and phases inferred from the first-layer factorization. We assume for now that the structure in amplitude and phase is independent, and thus we learn separate models for each set of variables. We first show how form-selective invariances may be learned from the log amplitudes, which we refer to as amplitude components. We then show how motion-selective invariances may be learned from the time derivative of phases, which we refer to as phase-shift components. Finally, we evaluate the ability of the model to represent form- and motion-selective invariances on a set of test image sequences.

4.1 Amplitude Components. The complex basis function model assumes independence between the subspace amplitudes and ignores dependencies among larger groups of units. A number of researchers have pointed out the dependencies of nearby linear filters among groups substantially larger than two, such as in the variances of nearby filters (Simoncelli, 1997; Schwartz & Simoncelli, 2001; Lyu & Simoncelli, 2009) or in the circular joint distributions of neighboring filters (Wegmann & Zetzsche, 1990). Extensions of ICA also model dependencies among cliques of filters, as in topographic ICA (Hyvärinen, Hoyer, & Inki, 2001), or among large groups of filters, as in independent subspace analysis (Hyvärinen & Hoyer, 2000).

To learn the joint structure among amplitudes in the first layer, we use another sparse coding model in the second layer (see Figure 1). However instead of modeling the amplitude values directly, we model the log amplitudes. This is similar to the approach of Karklin and Lewicki's (2005) density components model where here the amplitudes a play the role of the scale factors λ in their model. We model the logarithm of the amplitudes because it maps the highly skewed, nonnegative amplitude values to a more uniform (or approximately gaussian) distribution occupying the entire real domain. Note that transforming the marginal distributions of coefficients has been utilized by a number of image modeling researchers (Chen & Gopinath, 2000; Shan, Zhang, & Cottrell, 2007; Lyu & Simoncelli,

[2] The linear relationship between the time derivative of phase and motion is a different effect from the linearization of the joint distribution dependencies of phase derivatives.

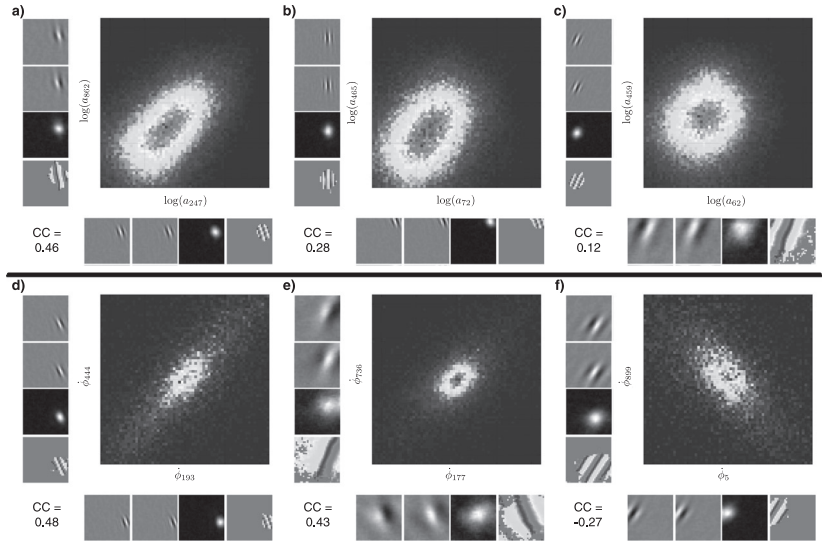

Figure 4: (a–c) Log amplitudes exhibit linear dependencies. Shown are the joint distribution of pairs of log amplitudes. Along each axis are shown the real, imaginary, amplitude, and phase plots for the corresponding complex basis function. There are clear linear correlations in the joint distributions of coefficients with basis functions overlapping in space (a: correlation coefficient = 0.46), nearby in space (b: cc = 0.28), and of different spatial frequency (c: cc = 0.12). (d–f) Phase derivatives exhibit linear dependencies. Shown are the joint distribution of pairs of phase derivatives. There are clear linear correlations in the joint distributions of coefficients with basis functions overlapping in space (d: high spatial frequency, correlation coefficient = 0.48; e: low spatial frequency, cc = 0.43), and at nonoverlapping spatial positions (f: cc = −0.27).

2009) and has an interesting relation to divisive normalization (Simoncelli, 1997; Schwartz & Simoncelli, 2001; Sinz & Bethge, 2008).

The logarithm of the amplitudes also has the property of linearizing dependencies between amplitudes, as demonstrated in Figures 4a to 4c. Here we show the joint distribution of log amplitudes for three different pairs of first-layer units. The amplitudes were generated by inferring the coefficients in the first layer for an ensemble of natural movie sequences. As one can see, the log amplitudes corresponding to basis functions at similar orientations and spatial positions show high linear correlations (see Figure 4a, correlation coefficient = 0.46). Coefficients with larger separations in space and scale show weaker linear correlations (see Figure 4b cc = 0.28, Figure 4c cc = 0.12). The presence of these correlations indicates clear dependencies among these variables that are well suited to be captured by a linear generative model.

The generative model for the log amplitudes is given by

$$\log a_{i(t)} = \gamma_i^0 + \sum_j B_{ij} v_{j(t)} + \rho_{i(t)}, \tag{4.1}$$

where γ_i^0 is a constant that sets the operating point for the linear model and $\rho_{i(t)}$ is additive gaussian noise with variance σ_a. The resulting prior on the amplitudes is then

$$P(a_{i(t)}|v_{(t)}) \propto e^{-\frac{1}{2\sigma_a^2}\left[\log a_{i(t)} - \gamma_i^0 - \sum_j B_{ij} v_{j(t)}\right]^2}, \tag{4.2}$$

where each column of B is an amplitude component basis function (in the space of the log amplitudes) that is multiplied by its respective scalar coefficient $v_{j(t)}$. The term σ_a corresponds to the noise variance in the log amplitude domain, which is small compared to the variation in the log amplitudes. We seek a small number of causes for the amplitude structure and expect that the representation of this structure will change slowly through time (it will be temporally stable). Therefore, we place a sparse and slow prior on the amplitude component coefficients:

$$P(v_{j(t)}|v_{j(t-1)}) \propto e^{-\lambda_v Sp_v(v_{j(t)}) - \beta_v Sl_v(v_{j(t)}, v_{j(t-1)})}. \tag{4.3}$$

For these simulations, we use $Sp_v(v_{j(t)}) = |v_{j(t)}|$, corresponding to a Laplacian prior and $Sl_v(v_{j(t)}, v_{j(t-1)}) = (v_{j(t)} - v_{j(t-1)})^2$.

The resulting negative log posterior or energy function for the amplitude portion of both the first- and second-layer model (ignoring an additive normalization constant) is given by

$$\begin{aligned} -\log P(a, v) = \quad & \sum_{t,i} \frac{1}{2\sigma_a^2}\left[\log a_{i(t)} - \gamma_i^0 - \sum_j B_{ij} v_{j(t)}\right]^2 + \\ & \lambda_v \sum_{j,t} Sp_v(v_{j(t)}) + \\ & \beta_v \sum_{j,t} Sl_v(v_{j(t)}, v_{j(t-1)}). \end{aligned} \tag{4.4}$$

Currently we infer the second-layer units by minimizing this function with respect to v only, holding the amplitudes a fixed to their values inferred using equation 3.7. The amplitude components of the second layer B are

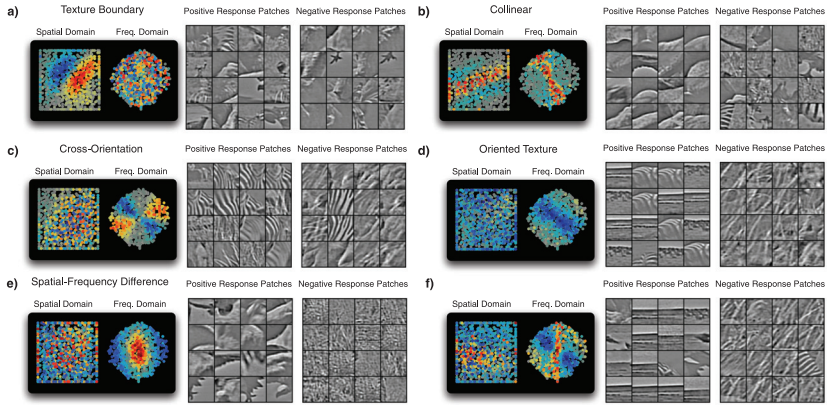

Figure 5: Structure of learned amplitude components. Each panel (a–f) illustrates the structure of a representative amplitude component from the population learned on natural movies. For each component, we provide two visualizations: (left) learned amplitude component weights and (right) exemplar image patches yielding large positive and negative responses. The learned weights (left) are visualized in the space of the first-layer basis functions (see the text). Image patches that produce a large positive or negative coefficient v (right) are selected from a large corpus of natural movies.

then learned by gradient descent on this same function, using the inferred values of a and v (see the appendix).

Some representative examples of amplitude components learned from natural movies are shown in Figure 5. Depicting what each amplitude component has learned is challenging because the weights live in the space of the first-layer units, and since the mapping is highly nonlinear, one cannot simply display the result in the image domain. We illustrate the forms of structure learned in the amplitude components by (1) depicting the weights in the space of the first-layer basis functions, (2) showing image patches that highly activate specific amplitude components, and (3) selecting exemplars from subpopulations that represent different types of amplitude component structure. For method 1, we depict the learned weights by utilizing the organization of the first-layer basis functions in the space and spatial-frequency domains, following the convention of Karklin & Lewicki (2005). Using the tiling of first-layer basis functions shown in Figure 2b, we illustrate an amplitude component j by coloring each dot corresponding to a first-layer unit, i, by its weight B_{ij}. Positive values of B_{ij} are mapped to shades of red, negative values are mapped to shades of blue, and values close to zero are mapped to gray.

We focus on prominent types of structure that emerge in the population of 625 learned amplitude components:

Texture boundary: A large number of amplitude components are selective for spatial texture boundaries. For example, the amplitude component depicted in Figure 5a has learned a texture boundary at roughly 45 degrees and localized at a particular position in space. In the space domain, the first-layer units with large positive weights and those with large negative weights show a clear separation in space and meet at a diagonal boundary. The organization of the weights for this function in the spatial frequency domain lacks clear structure. Therefore, this function is selective for the spatial pattern of image content but is invariant to the composition of the orientation and spatial frequency structure. Because the weights of the amplitude components are coupled to the logarithm of the amplitude coefficients, large negative weights will attenuate the corresponding amplitude coefficient toward zero, whereas large positive weights will amplify the amplitude coefficient toward a large positive value (for $v_j > 0$). This will effectively suppress the presence of image contrast or structure in one spatial region and enhance the contrast of image structure in the other region.

In the right portion of Figure 5a, we show 16 randomly selected image patches that produce a high positive response for this amplitude component and 16 randomly selected image patches that produce a high negative response. These image patches generally reflect the texture-boundary selectivity of this component. The insensitivity of this amplitude component to orientation and spatial frequency makes it invariant to the structure within a region of texture, yet the space domain selectivity makes it sensitive to the spatial envelope.

Collinear: Another prominent type of amplitude component groups together first-layer functions that are collinear and contiguous, and span a broad spectrum of spatial-frequencies, as seen in Figure 5b. This function has large positive weights to first-layer units that are spatially organized in a diagonal pattern at roughly 45 degrees. The preferred orientation of these first-layer units is collinear with this diagonal pattern. Also note that this function has large weights to basis functions that span low, medium, and high spatial frequencies. It has a broadband organization that integrates across spatial frequency. Image patches that produce a highly positive inferred coefficient all exhibit some collinear structure in the 45 degree direction. Image patches that produce a highly negative inferred coefficient are less similar since the negative weights are rather diffuse, but there is a clear lack of elongated contour structure of the preferred orientation and spatial location.

Cross-orientation: We also observe a large population of amplitude components that exhibit cross-orientation tuning, as seen in Figure 5c. Interestingly, the orientation difference is not perpendicular, or 90 degrees, but closer to 60 degrees. The spatial pattern also has a characteristic structure, with basis functions tuned to one orientation having a somewhat collinear organization and basis functions tuned to the 60 degrees offset orientation being more spatially homogeneous and lacking clear collinearity. Image

patches that produce large positive responses all have a predominant orientation structure at about −30 degrees and patches that produce large negative responses have image structure at about +30 degrees. Note that a number of aspects of the image structure have high variability, such as the exact position of prominent edges or the spatial frequency content of the edges.

In Figures 5d, 5e, and 5f, we show some additional learned amplitude components. In Figure 5d, we show an amplitude component that is orientation selective and has broad spatial tuning. Note the strong preference for textures across the entire image patch. In Figure 5e, we show an amplitude component that has antagonistic selectivity to spatial frequency. This component is differentially selective for low versus high frequencies and has a slight orientation selectivity. We also observe a number of amplitude components that are not as easily classified. In Figure 5f, we depict an amplitude component with an interesting differential tuning to a narrowly selective horizontal frequency structure and a broadly orientation tuned structure at about 45 degrees. We have observed such amplitude components on multiple runs of our algorithm and at varying spatial positions and orientations.

We illustrate the prominent subpopulations of amplitude components and their tiling properties in Figure 6. Each column corresponds to a different subpopulation, with exemplars of each shown by row. Note that each column shows only a small subsample for each of these subpopulations. Figure 6a shows five exemplar amplitude components that are selective for texture boundaries. These components span a range of spatial positions and orientations. The first three components are of the same boundary orientation and tile spatial position; the last two show evidence for the tiling in orientation. Figure 6b shows a similar tiling of space and orientation for components selective for an elongated edge structure. Figure 6c shows the range of structure in the cross-orientation selective components. While these components are broadly tuned spatially, they do show clear localization in space (the component in the first row is spatially localized in the lower left region of space). Note also the clear tendency for orientation opponency at roughly 60 degrees. Figure 6d shows the organization of components that have smooth variations in space and sometimes weak orientation tuning.

4.2 Phase-Shift Components. The dynamics of moving objects or observer motion over short epochs are encoded in the time rate of change of the entire population of phase variables. Local motion in the image domain that would otherwise produce nonlinear trajectories in the basis function coefficients $\dot{u}_i^{\mathcal{R}}(t)$ or $\dot{u}_i^{\mathcal{I}}(t)$ will now be linear in the corresponding phase derivative $\dot{\phi}_i = \phi_i(t) - \phi_i(t-1)$ (as demonstrated in Figure 3). However, complex natural motions will exhibit dependencies in the joint distribution of phase derivatives (dependencies among multiple phase derivative variables). Importantly for our use of a generative model in the second layer, the process

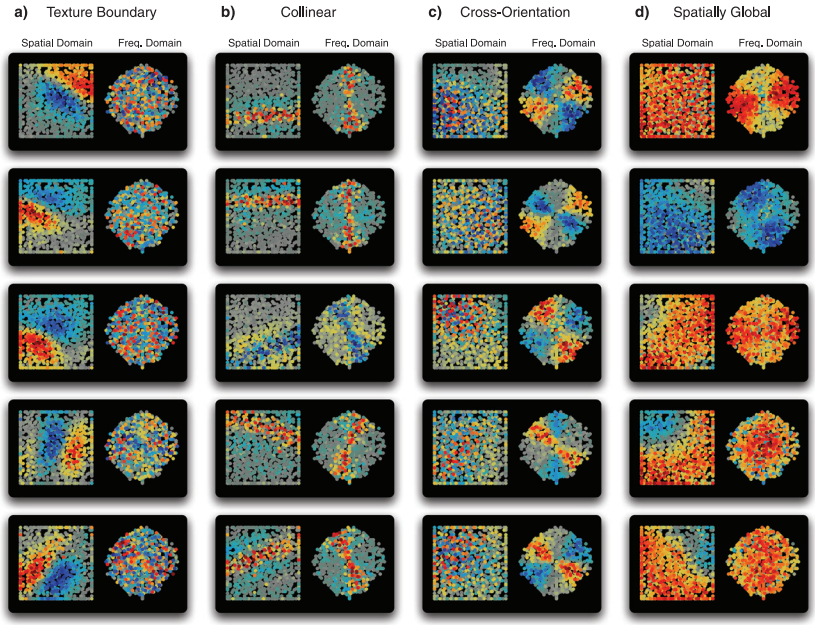

Figure 6: Amplitude component subpopulations. Each column illustrates a typical group of learned amplitude component functions. (a) Texture-boundary. (b) Collinear edge. (c) Cross-orientation. (d) Broad spatial tuning and broad orientation tuning. The texture-boundary selective components and the collinear edge components span a range of spatial positions and orientations. The cross-orientation components are more broadly tuned spatially but also span a range of orientations and spatial positions (the component in the first row is spatially localized to the lower-left region of space). The components in *d* are broadly tuned in space and orientation.

of factorization and taking the phase derivative produces a linearization of these dependencies in the joint distribution. We next demonstrate that the joint statistics between multiple phase derivatives show clear linear dependencies.

Figures 4d to 4f show the joint phase derivative distribution for three pairs of complex basis functions. These distributions are produced by inferring the first-layer variables for an ensemble of natural movies. Complex basis functions with similar orientation, position, and spatial frequency show high correlations (Figure 4d, cc $= 0.48$). Functions with different orientation also show high correlations (Figure 4e, cc $= 0.43$), likely due to coherent motion of textures in natural movies. Functions with nearly no spatial overlap also show correlation of significant magnitude (Figure 4f,

cc = −0.27). Note that the negative correlation (versus positive) is arbitrary and is determined by the handedness of the complex basis functions, which is symmetric in the model formulation and determined only by the initial condition before learning. Just as with the observed log-amplitude correlations, the presence of linear correlations of the phase derivatives indicates clear dependencies among these variables that are well suited to be learned in a linear generative model. Therefore, the process of factorization and taking the phase derivative produces a linearization of the dependencies in the joint phase distribution.

The generative model for the time derivative of the phase variables is given by

$$\dot{\phi}_{i(t)} = \sum_k D_{ik}\, w_{k(t)} + \nu_{i(t)}, \tag{4.5}$$

where D is the basis function matrix specifying how the high-level variables w_k influence the phase shifts $\dot{\phi}_i$. The additive noise term, ν_i, represents uncertainty or noise in the estimate of the phase time-rate of change. The generative model is shown schematically in Figure 1. As before (see equation 4.3), we impose a sparse and slow prior on the second-layer coefficients w_k:

$$P(w_{k(t)}|w_{k(t-1)}) \propto e^{-\lambda_w Sp_w(w_{k(t)}) - \beta_w Sl_w(w_{k(t)}, w_{k(t-1)})}, \tag{4.6}$$

with the sparse cost function in this case given by $Sp_w(w_{k(t)}) = \log(1 + w_{k(t)}^2/\sigma^2)$ and the slowness penalty given by $Sl_w(w_{k(t)}, w_{k(t-1)}) = (w_{k(t)} - w_{k(t-1)})^2$. Slowness in this case corresponds to our expectation that motions in the visual world are persistent through time; for example, objects moving rightward tend to continue to move rightward over time (a direct consequence of physical momentum).

The uncertainty of the generated phase shifts is given by a von Mises distribution: $p(\nu_i) \propto \exp(\kappa \cos(\nu_i))$, which is a univariate generalization of the gaussian distribution to an angular variable. The resulting conditional distribution of the first-order time derivative of the phase, given the coefficients w, is

$$P(\dot{\phi}_{i(t)}|w_{(t)}) \propto e^{\kappa \cos(\dot{\phi}_{i(t)} - [Dw_{(t)}]_i)}. \tag{4.7}$$

The notation, $[Dw_{(t)}]_i$ indicates the ith row of the matrix product.

When the priors on the first- and second-layer variables are combined, the resulting negative log posterior for the phase portion of the model

(ignoring an additive normalization constant) is given by

$$-\log P(\phi, w) = -\sum_{t} \sum_{i \in \{a_i(t) > \epsilon\}} \kappa \cos(\dot{\phi}_i(t) - [D\, w(t)]_i) + \lambda_w \sum_{k,t} Sp_w(w_{k(t)}) + \beta_w \sum_{k,t} Sl_w(w_{k(t)}, w_{k(t-1)}). \quad (4.8)$$

Note that because the phase of a complex variable with amplitude close to zero is undefined, we exclude $\phi_{i(t)}$ where either $a_{i(t)}$ or $a_{i(t-1)}$ is less than a small constant, ϵ.

As with the amplitude model, we infer the second-layer units by minimizing equation 4.8 with respect to the w only, holding the first-layer phases ϕ to their values inferred from equation 3.7. The phase-shift components D are then learned by gradient descent on equation 4.8, using the inferred values of ϕ and w (see the appendix).

We next describe the properties of the phase-shift components D learned from natural movies. We attempt to convey the structure of these components by examining their weights, estimating motion vectors from generated transformations, and illustrating the tiling properties of the population.

In Figure 7, we depict six representative components. The phase shift component in Figure 7a is perhaps easiest to understand because it corresponds to vertical translation throughout the entire image patch. In the space domain, there are large weight values distributed over all positions, while in the frequency domain, there is a ramp in the vertical direction. This latter structure is due to the fact that coherent translations in the image domain are produced by phase shifts that are proportional to spatial frequency and also depend on the alignment between a unit's orientation and the direction of motion. Basis functions represented by mirror symmetric points in the 2D Fourier plane will precess in opposite directions for a positive change in their coefficients' phases.

We also visualize the effect of the image domain transformation by estimating local motion vectors from image domain transformations produced by a phase-shift component (see the appendix for details). Using this method, we display the spatial aspects of the induced transformation on the right in Figure 7a, which illustrates the global vertical motion induced by this phase shift component.

We have found that generating movies using each phase-shift component is highly instructive for discerning what the component has learned. To produce such a movie, we select a single static image patch and infer the amplitude and phase coefficients for the first layer. Given these amplitude and phase coefficients, we then produce a phase shift by holding a

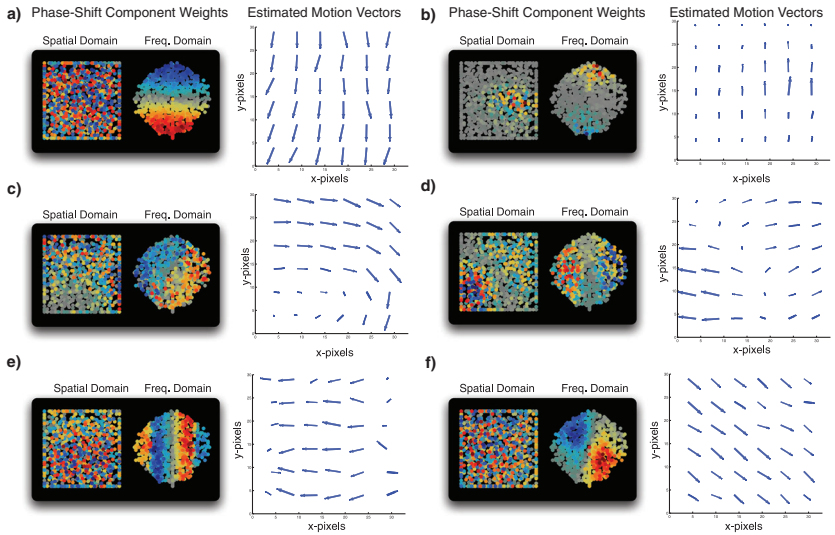

Figure 7: Structure of learned phase-shift components. Each panel illustrates the structure of a representative phase-shift component from the population learned on natural movies. For each subpanel, we provide two illustrations: (left) learned phase-shift component weights and (right) motion vectors estimated from generated transformations. The learned weights are visualized in the space of the first-layer basis functions using the same convention as the amplitude components in Figure 5. Motion vectors indicate the image domain motion produced by a positive contribution of the corresponding phase-shift component (see the appendix). Each component produces a unique transformation in the image domain. The components are (a) global vertical translation, (b) local vertical translation, (c) rotation, (d) dilation, (e) temporal-aliased structure, and (f) complex translation. Movies showing generated image transformations for each of these components can be found online at http : //www.vimeo.com/album/1624584.

second-layer unit w_k at a constant value over a number of frames, causing a certain pattern of phase precession. (The resulting movies may be viewed online at http://www.vimeo.com/album/1624584.) In these movies, it is perceptually evident that each of the animated image patches undergoes a similar transformation, even though their spatial structure is quite different. This is perhaps the strongest evidence that these learned components reflect motion-selective invariances.

In addition to global phase-shift components, the model also learns spatially localized phase-shift components as in Figure 7b. This component produces a vertical translation localized to a spatial region just to the right of the center of the image patch, which is also evident in the estimated

motion vectors. The generated motion is most prevalent in image patches that have horizontally oriented image content in the relevant image region. Note that the image regions outside this zone are left unaltered by this component.

While translation operators comprise the majority of the learned components, a large number exhibit other forms of interesting structure. The component in Figure 7c, for example, produces a rotational warping in the image domain around a point that appears to be in the lower left of the image patch. This is evident in the weight pattern for this component as the magnitude of the weights in the lower left of the image patch are quite small and the estimated motion vectors spiral around this point. The component in Figure 7d produces a type of dilation or expansion in the image domain. The left and right portions of the image patch exhibit motion in opposite directions. When the right portion undergoes a translation to the right, the left portion translates to the left. As the induced transformation changes sign, the right portion of the image translates leftward, and the left portion translates rightward. The weight pattern also suggests a particular spatial selectivity for this component with the upper left and lower right of the image patch remaining static for the transformation, which is reflected in the estimated motion vectors.

The model also learns a number of phase-shift components that we either did not expect or for which we have not found precise interpretations. The phase-shift component depicted in Figure 7e we believe is related to temporally aliased motion structure. The weight pattern is unexpected as it implies that high-frequency phases should advance in the opposite direction as low-frequency first-layer phases. Indeed, when we animate image patches using this component, the motion that is produced appears to be two transparent layers distinguished by their spatial frequency and moving in opposite directions (leftward or rightward). Upon inspecting movie sequences where these components are used, it appears that this structure is due to temporal aliasing of motion in the movies, which occurs in the horizontal direction. This is a consequence of the presence of many movie sequences in the data set that contain fast horizontal motion in the background as animals are being tracked in the foreground. Therefore this structure is likely due to an insufficient temporal sampling rate for the observed motions. The component in Figure 7f produces a transformation that appears to be a translation. However, the weight pattern for this transformation has a peculiar structure in the spatial frequency domain with two wedge shapes not aligned with the radial direction. Although we do not have a concise interpretation of this learned pattern, it does appear significantly often in the learned population.

We illustrate the variety of the learned phase-shift components within each class in Figure 8. Figure 8a illustrates different directions of motion within the population that are selective for global translation (spanning directions $+45^\circ$, $+90^\circ$, $+135^\circ$, $+180^\circ$). Figure 8b shows the range of spatial

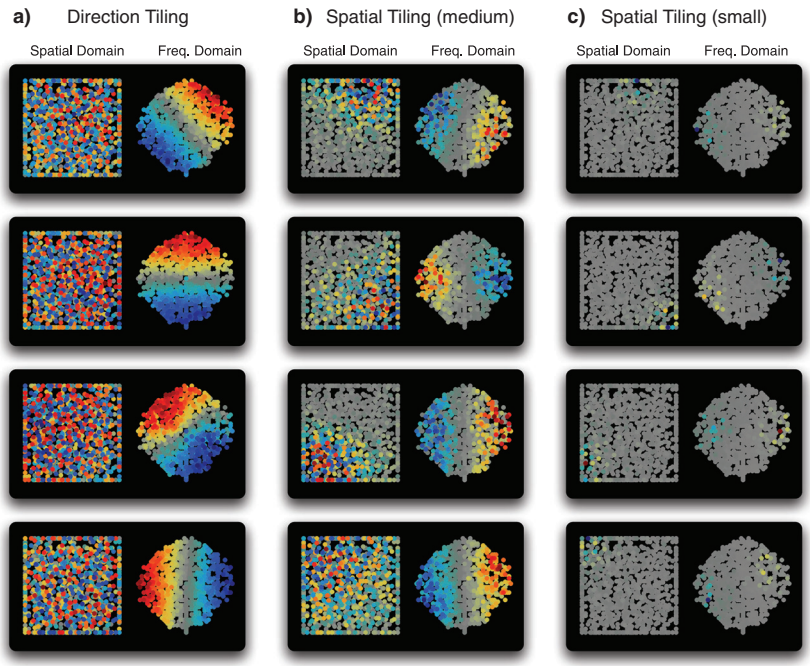

Figure 8: Phase-shift component subpopulations. Each column depicts four exemplar phase-shift components. (a) Components selective for global translation are selective for different directions of motion. (b) Components of medium spatial extent with selectivity for horizontal motion tile spatial position, from top to bottom: top right, lower right, lower left, top left. (c) Components of small spatial extent selective for horizontal motion tile space.

positions of components selective for motion of the same direction, from top to bottom: upper right, lower right, lower left, and upper left. Figure 8b shows part of the spatial tiling for components of medium spatial extent, and Figure 8c shows the same tiling property for components with small spatial extent. Note that only a small fraction of the entire population for each class is shown here. The variety of spatial sizes, positions, and directions in the phase-shift components can be seen as forming a multiscale, compositional code for describing image translation over arbitrarily shaped regions, thus providing a possible mechanism for motion segmentation.

4.3 Testing Form- and Motion-Selective Invariance. To test form- and motion-selective invariance in our model, we generated movies that vary either the image form or the image motion. We used these movies to assess the degree of variation among different layers in response to changing

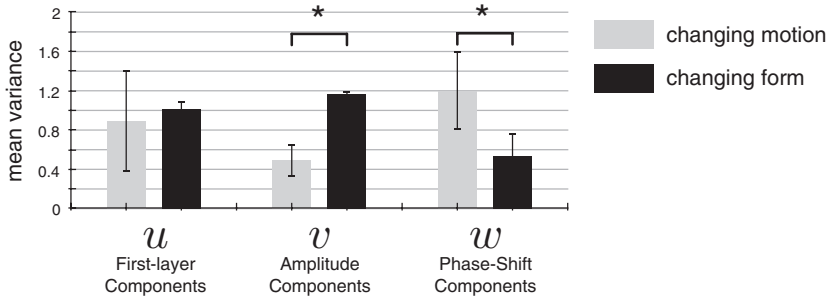

Figure 9: Testing form- and motion-selective invariance. The degree of variation in each set of variables in the model (u, v, w) was measured in response to changes in image motion (gray) or to changes in image form (black).

either the image form and keeping the image motion fixed or changing the image motion and keeping the image form fixed. We compare the amplitude-component coefficients v to the phase-shift component coefficients w and use the first-layer coefficients u^{R} as a control comparison (see the appendix). Figure 9 shows the results of this test. As expected, the first-layer coefficients show no significant difference (determined by t-test, $p = 0.12$) in mean variance when changing the image motion or the image form (the variation in these coefficients is due equally to changes in motion and changes in image content). However, the amplitude components show higher variance as image form changes and lower variance as image motion changes ($p = 1.3 \times 10^{-13}$). This indicates that the amplitude components are selective for image form and invariant to image motion. Conversely, the phase-shift components show low variance across changing image form and show higher variance across changing the image motion ($p = 7.2 \times 10^{-5}$). Therefore, these components are selective for image motion and invariant to image form.

5 Discussion

We have proposed a hierarchical model of image representation and have shown that it is capable of learning rich representations of form and motion contained in natural image sequences. In this section, we discuss the forms of structure learned in the first and second layers, the relationship of our model to other models of visual processing, the importance of factorization and its implications for models for form and motion processing in visual cortex, and the limitations of the particular model we have proposed.

5.1 Complex Basis Function Representation. The complex basis function representation used in the first layer of our model shares similarities

with other natural image models. A number of investigators have extended the sparse coding model to include dependencies among statistically dependent filters by either imposing or learning groupings among features, such as independent subspace analysis (Hyvärinen & Hoyer, 2000), topographic-ICA (Hyvärinen et al., 2001), and the mcRBM model (Ranzato & Hinton, 2010). Each of these models imposes certain restrictions on the dependencies, or groupings, that are allowed between filters. The complex basis function model used here may be viewed as the most restrictive along this axis as it allows only pairwise dependencies. However, the main purpose served by the complex representation is simply to allow for an explicit factorization into amplitude and phase from which higher-order groupings can be learned by the second layer. In this sense, the complex pairing simply mediates factorization; it is not an end goal of the representation in itself.

Slowness, or temporal persistence, has also been a recurring theme in the study of natural image statistics. Models taking advantage of slowness also find similar image structure as our first-layer model: for example Wiskott and Sejnowski (2002), Einhauser et al. (2002), Berkes & Wiskott (2005) each find groupings of oriented filters that span subspaces invariant to phase. In our model, we have extended the slowness constraints to the second-layer representations of form and motion.

5.2 Learned Amplitude Components. The higher-order structure learned by the amplitude components is qualitatively similar to that learned by other models of amplitude or variance modulation of image filters. In particular, because of its similar mathematical form, Karklin and Lewicki's density component model exhibits similar learned structure to that shown here (Karklin & Lewicki, 2003a). This is not surprising as the structure is dictated by natural images. However, a novel class of second-layer functions emerges in our model that is not seen in the published work on the density components model (Karklin & Lewicki, 2003a, 2003b, 2005) or related models (Schwartz et al., 2006). This class of functions represents elongated, collinear edge structure in the image domain (see Figure 5b). It has not been shown if more recent models, similar to the density components model, do in fact capture this elongated edge structure (Karklin & Lewicki, 2008). There is evidence that the multilayer ICA model (Hyvärinen, Gutmann, & Hoyer, 2005) captures multiscale edges, but the diversity of structure learned by this model is limited compared to the amplitude components presented here.

The ability to learn elongated, collinear edge structure is likely due to the local translation invariance that is provided by the amplitudes of first-layer units. The multiplicity of positions and orientations that an extended contour takes on within a 32×32 pixel image patch would make it impractical for a set of basis functions to completely tile this space. This combinatorial explosion is mitigated in our model by the phase shifting that occurs in the first-layer units. The result is that a given amplitude pattern specified by

the second layer can actually correspond to a wide variety of contours with slightly different positions, orientations, or relative phase offsets along the contour. By contrast, in the density components model, the value of each first-layer unit will be highly sensitive to these variations, making it more difficult for the second layer to learn the underlying form invariance of elongated contours. This difference also highlights the importance of factorization in our model: since variations in spatial position can be explained away by phase, the amplitude is free to carry the underlying invariant information about the presence of edges. Without this explaining away that occurs in the first layer, form and spatial variation (motion) would still be entangled, and the underlying form invariances would be less visible to the second layer.

One complication in our model that is introduced by modeling the log amplitudes is that amplitude values close to zero get mapped to large negative numbers. Although the difference between $a_i = 0.01$ and $a_i = 0.000001$ may be insignificant, it will become highly significant in the log domain and penalized due to the l_2 distance in equation 4.2. The sensible thing to do in this case would be to set $a_i = 0$, but the logarithm function does not allow this. One possibility is to use a thresholded log function that takes the logarithm only for values above a threshold, setting values below threshold to zero. This or other solutions should be explored in future work.

Finally, it is important to note that other types of form-related information are not captured by the amplitude components. Namely, the absolute phase in the first-layer units carries important information about relative spatial relationships among image features. Currently there is no prior over the absolute phase (only the phase derivative). As a consequence, it is still not possible to generate realistic images from the model. How to model the structure in absolute phase is an important and difficult open problem. Some recent progress has been made on this issue through the development of multivariate models of phase dependencies (Cadieu & Koepsell, 2010a, 2010b), and incorporating these ideas into the model is a goal of ongoing work.

5.3 Learned Phase-Shift Components. The structure learned by the phase-shift components captures a rich variety of transformations that occur in natural image sequences. This is in contrast to the first-layer units or other single-layer models of time-varying images such as Olshausen (2002), which cannot explicitly represent motion because their units are highly localized in space, orientation, and spatial frequency and are thus not invariant to these properties of the moving structure. The second layer of our model overcomes this problem by integrating across phase derivatives in the first layer that are consistent with a certain type of image transformation, thus forming an explicit representation of motion that is invariant to local orientation and spatial frequency structure.

The use of phase to represent local image shifts is also not unique to our model and has been used previously by other investigators to compute motion (Fleet & Jepson, 1990; Magarey & Kingsbury, 1998). One of the advantages of phase is that it is insensitive to contrast variations. Here we utilize in addition the fact that phase linearizes curved trajectories in coefficient space and thus allows the second layer to capture the higher-order motion structure via a simple linear generative model (see Figures 3 and 4). The insensitivity of phase to contrast variations is also key to the motion-selective invariance achieved by the phase-shift components. Another class of models of biological motion processing (Simoncelli & Heeger, 1998; Rust, Mante, Simoncelli, & Movshon, 2006) implicitly removes form information by dividing out the local image contrast. This divisive normalization operation is analogous to the separation of amplitude and phase in the first layer of our model. We can make this relationship explicit by expressing the time derivative of the phase in terms of quadrature-pair simple cell responses, u_c and u_s,

$$\frac{d}{dt}\phi(t) = \frac{d}{dt}\arctan\left(\frac{u_s}{u_c}\right) = \frac{\dot{u}_s u_c - u_s \dot{u}_c}{u_c^2 + u_s^2} = \frac{u_s u_c^{t-1} - u_s^{t-1} u_c}{u_c^2 + u_s^2}, \tag{5.1}$$

where the last equality is achieved by approximating the time derivative with a first-order difference. This relationship, described by Simoncelli (1993), shows an alternative way to compute the phase-shift variables in our model in terms of variables that are readily available in divisive normalization models (Simoncelli & Heeger, 1998; Rust et al., 2006).

The structure that emerges within the phase-shift components reflects the dynamics and structure contained in natural movies. The diversity of this structure has not been addressed by previous models of motion processing (Nowlan & Sejnowski, 1995; Zhang et al., 1993; Grimes & Rao, 2005; Rolls & Stringer, 2007). One of the most important questions in computing motion is how to build a complete representation that tiles the joint domain of space (spatial position) and motion (speed and direction). The model in Simoncelli and Heeger (1998), for example, proposes specific weight patterns among spatiotemporal filters that are hand-tuned to reproduce physiological data, but the question of how to design an entire population of such units to encode the complex motion that actually occurs in dynamic natural scenes is unaddressed. These details are precisely what is learned by our model: the majority of phase-shift components correspond to image translation operators that are localized within different regions of the space domain and extend over different spatial scales, allowing for the complex segmentation of motion in time-varying natural images. Previous models of biological motion processing (Simoncelli & Heeger, 1998; Rust et al., 2006) have not specified over what spatial extent motion should be computed, and indeed it would seem difficult to determine this from first principles. Here the

solution is learned by the weight patterns in the second layer of our model to match the statistics of natural image motion.

5.4 Factorization in Visual Models. Factorization plays a key role in our model and theoretically extends back to early proposals of visual layers (Barrow & Tenenbaum, 1978). More recent models have shown how factorization can be used to separate style and content in image data (Tenenbaum & Freeman, 2000), and related approaches have been utilized in computer vision problems (Tomasi & Kanade, 1992; Koenderink & Van Doorn, 1997). Importantly, factorization is not a feedforward computation. In factoring a number, for example, one is not free to choose each factor independently; the computation of one factor necessarily influences the other. In our model, this interaction occurs in the factorization of amplitude (which signals the presence of edge structure) and phase (which encodes relative position) in the first-layer units. The slow and sparse prior on amplitudes encourages representations in which image features persist over time. The phase must then accommodate to account for changes in the image data. The inference of one parameter (presence of a feature) depends on the other (position), and vice versa. Note that the slowness prior on the amplitude plays a role akin to the constant brightness assumption in motion estimation (Horn & Schunck, 1981), and the phase plays the role of the shift parameter.

Factorization has played an important role in other models of natural images. In particular, the bilinear model of Grimes and Rao (2005) factors the position of a filter component apart from the amplitude of that component. However, in that model, the position variable is shared globally among all features in the representation and is thus more restrictive. Our less restrictive approach of having a position (phase) variable for each feature enables us to learn the structure of motion from unsupervised exposure to natural images, as opposed to using controlled artificial motions as in Grimes and Rao (2005). The model in Berkes et al. (2009) is closely related to the first layer of our model and has some specific advantages. This model factors the image filter coefficients into a binary presence variable and an appearance component. The learned subspaces can be larger than two and also group together units of similar spatial position, orientation, and scale. However, our model uses the factorization in the first layer merely as a staging ground for learning higher-order structure in the second layer. It seems likely that structure similar to that found in our second-layer model could be learned from the first-layer representation in Berkes et al. (2009).

5.5 Neurobiological Implications. The development of this model was motivated by considerations about the structure contained in natural images and how to extract invariances rather than by the desire to explain or account for specific neurobiological phenomena. However, to the extent that the visual system has been adapted to the structure of dynamic natural images, we may expect to find a relationship between the types

of form and motion representations discovered by our model and neural representations found in the brain. Indeed, the separation of form and motion processing in our model would seem to mirror the what and where streams found in visual cortex. There is an important distinction, however, between the manner in which form and motion are computed in our model and the standard models of form and motion processing in visual cortex (Simoncelli & Heeger, 1998; Serre et al., 2007). Namely, we propose that these properties are extracted using a factorization process in which the two computations interact rather than being computed independently. It may be possible to test if this is the case by generating stimuli that violate the constancy assumption in form and look at how the representations of form and motion are affected. Certain psychophysical phenomena such as motion silencing (Suchow & Alvarez, 2011) seem to be consistent with the idea that the assumption of object constancy overrides local changes in an object when it is moving.

An obvious question that arises is how this model would be implemented in terms of neurobiological substrates and mechanisms. At first glance, the representation of complex numbers and quantities such as phase may seem rather implausible from what we know of the existing physiological data. However, as we have seen from the discussion, there are ways to represent phase and phase derivatives that are consistent with existing divisive normalization models of simple cells. The first-layer units of our model would thus seem most directly related to complex cells and normalized simple cells in V1. The second-layer units could possibly be instantiated at higher levels of representation such as V2 (for form) or MT (for motion). The learned amplitude and phase-shift components could be used as a basis for exploring representations in these higher-level areas, either by regressing the recorded activity of neurons against the activity of second-layer units in the model in order to see how well they predict neural activity or by generating stimuli from the model and examining how visual neurons respond to individual amplitude or phase-shift components. Regardless of whether the model has any validity from a neurobiological standpoint, it does offer a valid tool to explore visual representations in the brain because it provides a parameterized description of higher-order structure in dynamic natural images.

5.6 Caveats. A number of simplifying assumptions are built into our model that may limit its scope. For example, the first-layer basis functions are only a function of space and not time. This is an obvious discrepancy with responses commonly observed in primary visual cortex, where neurons are known to have nonseparable temporal and spatial receptive fields (DeAngelis, Ghose, Ohzawa, & Freeman, 1999). The data we use for training do not include color, binocularity, or natural fixational eye movements. These aspects of natural vision likely play important roles in determining the structure in intermediate visual cortex. In addition, the inference

procedure we use to learn model parameters is noncausal. It is unknown how constraining the inference and learning procedure to be causal would affect the results. Another modeling aspect we have not explored is the level of overcompleteness in the model, which has been shown to drastically affect the qualitative structure that is explicitly captured by one-layer sparse coding models (Olshausen, Cadieu, & Warland, 2009).

While the second-layer model captures dependencies in amplitude and phase shift, it ignores dependencies in instantaneous phase. In separate work, we have proposed statistical models that may be relevant for capturing these unmodeled dependencies (Cadieu & Koepsell, 2010a, 2010b), and it will be important to include these forms of structure in order to develop a complete model of higher-order form and motion.

A valid concern about the structure imposed by our model is the pairing of first-layer basis functions. Through learning on natural movies, the basis functions become quadrature pairs. The hypothesis that neurons in primary visual cortex are paired in quadrature was entertained by Pollen and Ronner (1981), but little evidence was found to support the hypothesis. In light of these negative results, we do not advocate an explicit pairing of neurons in primary visual cortex, but rather subscribe to models that specify arbitrary pairings of filters (Hyvärinen & Hoyer, 2000; Berkes et al., 2009; Karklin & Lewicki, 2008). The pairings we have used are a simplification that enabled us to tractably arrive at localized amplitude components with dynamic phase variables. An important aspect of the models that learn this grouping structure is that even though a multidimensional subspace is learned, the dynamics within the subspace during inference of moving stimuli often follows a low-dimensional trajectory, usually 1D (Berkes et al., 2009; Culpepper & Olshausen, 2009). This indicates that our approximation of these trajectories with a one-dimensional phase may be consistent when the dimensionality of the subspaces is learned, but the proper way to learn and represent these trajectories is still an open problem.

Appendix: Simulation Details

A.1 Learning and Inference. We seek to learn the parameters for the basis functions, A, B, and D, in both layers from image sequences. We used a variational learning algorithm to adapt the basis functions in both layers. First we infer the maximum a posteriori (MAP) estimate of the variables a, ϕ, v, and w for the current values of the basis functions. Given the MAP estimate of these variables, we then perform a gradient update on the basis functions. The two steps are iterated until convergence.

To infer coefficients in the first hidden layers, we perform gradient descent with respect to the coefficients of the cost function (see equation 3.7). The resulting gradients for the amplitudes and phases in the first layer are

given by

$$\Delta a_{i(t)} \propto \Re\{b_{i(t)}\} - \lambda_a S p'(a_{i(t)}) - \beta_a S l'(a_{i(t)}, a_{i(t-1)}), \tag{A.1}$$

$$\Delta \phi_{i(t)} \propto \Im\{b_{i(t)}\}\, a_{i(t)}, \tag{A.2}$$

with $b_{i(t)} = \frac{1}{\sigma_N^2} e^{-j\phi_{i(t)}} \sum_x A_{i(x)} \left[I_{(x,t)} - \sum_i \Re\{z^*_{i(t)} A_{i(x)}\} \right]$. $\Im\{.\}$ denotes the imaginary part. Note that here we consider inference only in the first layer independent of the second layer.

The gradients for the second-layer coefficients v_j and w_k are given by

$$\Delta v_{j(t)} \propto \frac{1}{\sigma_a^2} \sum_i \left[\log a_{i(t)} - \gamma_i^0 - [Bv_{(t)}]_i \right] B_{ij} - \lambda_v S p'_v(v_{j(t)}) \tag{A.3}$$

and

$$\Delta w_{k(t)} \propto \sum_{i \in \{a_{i(t)} > 0\}} \kappa \sin(\dot{\phi}_i - [Dw_{(t)}]_i)\, D_{ik} - \lambda_w S p'_w(w_{k(t)}) - \beta_w S l'(w_{k(t)}, w_{k(t-1)}). \tag{A.4}$$

The learning rule for the first-layer basis functions is given by the gradient of equation 3.7 with respect to $A_{i(x)}$, using the MAP estimates of the complex coefficients:

$$\Delta A_{i(x)} \propto \frac{1}{\sigma_N^2} \sum_t \left[I_{(x,t)} - \sum_j \Re\{z^*_{j(t)} A_{j(x)}\} \right] z_{i(t)}. \tag{A.5}$$

The learning rule for the amplitude components is given by the gradient of equation 4.4 with respect to B_{ij}, using the MAP estimates of the values of a and v:

$$\Delta B_{ij} \propto \frac{1}{\sigma_a^2} \sum_t \left[\log a_{i(t)} - \gamma_i^0 - [Bv_{(t)}]_i \right] v_{j(t)}. \tag{A.6}$$

The learning rule for the second-layer basis functions ΔD_{ik} is given by the gradient of equation 4.8 with respect to D_{ik}, using the MAP estimates of the values of ϕ and w:

$$\Delta D_{ik} \propto \kappa \sum_{t \in a_{i(t)} > 0} \sin(\dot{\phi}_i - [Dw_{(t)}]_i)\, w_{k(t)}. \tag{A.7}$$

After each gradient update, the length of each basis function is normalized to have unit length (l_2 norm). We also found that convergence improved when we orthogonalized the real and imaginary parts of each complex basis function using the Gram-Schmidt process after each update.

A.2 Data Sets and Simulation Details. We used natural scene data sets for training the model. For our experiments using natural movies, we used natural image sequences obtained from Hans van Hateren's repository, available at https://github.com/cadieu/twolayer . The movies were spatially low-pass-filtered and whitened as described previously (Olshausen & Field, 1997). No whitening in time was performed because the hierarchical model will learn the temporal structure. The movies consisted of footage of animals in grasslands along rivers and streams. They contain a variety of motions due to the movements of animals in the scene, camera motion, tracking (which introduces background motion), and motion borders due to occlusion.

We use stochastic variational learning to train the basis functions in both the first and second layers. In this learning algorithm, we estimate MAP estimates of the latent variables on small batches of data and then descend the gradient on the basis functions given these MAP estimates. We repeat this procedure until convergence. We trained on 32×32 pixel image patches using 1024 complex basis functions in the first layer initialized to random values, and 625 basis functions for both the B and D functions, also initialized to random values. In the initial learning phase, we first trained the first-layer model using only the terms in equation 3.7 to infer the a and ϕ variables. We estimated MAP values on randomly selected movie sequences of 128 frames. We determined that the first layer had converged after 240,000 iterations and annealed the learning rate. We began training the second-layer basis functions B and D only after the first layer had converged. The second-layer bases were trained on the MAP estimates of the first layer. We also performed stochastic gradient descent on the second-layer basis functions with batches of 128 movies frames and repeated for 180,000 iterations. We used Matlab for the code implementation and the Jacket GPU interface by Accelereyes. A version of this code can be found online at: https://github.com/cadieu/twolayer (this code also supports additional priors not discussed here and implements a patch-based whitening procedure instead of the frame-based procedure described here). Ideally, we should adapt both the first layer basis functions and the second layer functions to reach a global optimum, but here we make an approximation and consider the first layer and second layer independently. We have run the algorithm multiple times and have observed qualitatively similar results on each run.

A.3 Selecting Highly Active Amplitude Component Images. We demonstrate the visual selectivity of the amplitude components by showing

images that produce responses of high magnitude in Figure 5. We attain these responses by performing inference on randomly selected image sequences selected from the corpus of natural movies. Following the procedure we use for learning, we first infer MAP estimates of the first-layer amplitudes and phases. We then infer the MAP estimate of the amplitude component coefficients v. Because the magnitude of the v_j coefficients is influenced by the contrast of the image sequence and we are more interested in pattern selectivity than contrast modulation, we perform a soft normalization on the v_j MAP estimates to rank images for visualization of a specific amplitude component:

$$\bar{v}_{j(t)} = \frac{v_{j(t)}}{\sqrt{.1 + \sum_{j'} v^2_{j'(t)}}}. \tag{A.8}$$

Given the normalized values, $\bar{v}_j$, of a large number of image sequences, we select the eight most positively active image patches and the bottom eight most negatively activity image patches to visualize a specific amplitude component (see Figure 5, top right, in each panel).

A.4 Estimating Motion Vectors from Image Transformations. We estimated local motion vectors on image domain transformations produced by specific phase-shift components (see Figure 7, top right in each panel). To estimate localized motion vectors, we used a grid pattern containing broadband frequency content across the entire visual field (this pattern can be seen in Figure 7 bottom, middle row, last image). Using this image pattern we followed the procedure to generate image domain manipulations for a specific phase-shift component with positive coefficient $w^0_k = 4\pi/5$ (described in section A.5). For each motion vector, we selected a region of the original image patch and computed the mean-squared error between this patch and corresponding regions in the image patch generated from the phase-shift component. We computed this error for motion vector offsets between 0 and 4 pixels in .25 pixel increments in both the horizontal and vertical directions. We used bilinear interpolation for subpixel motion vectors. For each region, we selected the motion vector that minimized the mean-squared error. Repeating this over a grid of reference image patches produced a grid of motion vector estimates as seen in the top right of panels in Figure 7.

A.5 Generating Phase-Shift Component Manipulations. In order to show the influence of the generative phase-shift components, we manipulate the inferred phase patterns of specific image patches according to a phase-shift component of interest. The resulting image manipulations are shown as movies online at http://www.vimeo.com/album/1624584. To create these manipulations, we select an image patch and infer the MAP

estimates of amplitude $\hat{\mathbf{a}}$ and phase $\hat{\phi}$ for this image patch. We then add a multiple of the phase-shift component vector to the phase vector $\hat{\phi}$. The phase-shift component vector is thus the phase difference between these two phase patterns. The component-wise addition is thus

$$\hat{\phi}_i + D_{ik} w_k^0 = \tilde{\phi}_{ik}. \tag{A.9}$$

where k indexes the phase-shift component we are examining. The magnitude of w_k^0 is determined such that the range of the phase difference reaches a value: $\max_i D_{ik} |w_k^0| = 4\pi/5$, and the sign is either $+w^0$ or $-w^0$. This gives two new phase pattens, $\tilde{\phi}^+$, and $\tilde{\phi}^-$. We then use each of these phase patterns to generate an image (using the original MAP estimates of the amplitudes). These images demonstrate the influence of a positive and negative addition of a phase-shift component to an image. In the corresponding movies provided online at http://www.vimeo.com/album/1624584 , we smoothly vary the value of w_k^0 through its maximum extent to create a smooth image domain manipulation between the extremes.

A.6 Testing Form- and Motion-Selective Invariance. To test the selectivity and invariance of the inferred variables in our model, we constructed movie sequences by generating translation sequences of an image patch over an image taken from our training data set. We generated 20 smooth translation trajectories (producing different motions) and selected 20 different images (producing different form information). We combined all combinations of trajectories and images, giving 400 movie sequences. We then inferred the latent variables in our model for each sequence. In Figure 9, we compare the first-layer variables, $u^{\mathbf{R}}$, the amplitude components, v, and the phase-shift components, w. For each coefficient type, we normalized the total variance over all 400 movies sequences, all variables, and all time points within the sequences to 1. To determine the variation in each variable type due to either changing motions or changing the image, we computed the variance of each variable at each time point within the sequence. The variance was calculated either across the different images (to determine selectivity to the transformation and invariance to the image) or across the different trajectories (to determine selectivity to the image and invariance to the motion). For each variable type in the comparison, we then averaged the computed variances over individual variables and over the time points within the sequences. This produced a distribution of 20 values for each coefficient type by computing the variance across either images or transformations. The mean of these distributions is plotted in Figure 9, and the standard deviation is shown as error bars around the mean. Significance was determined with a standard t-test. In summary, the results indicate that the variation in the first-layer coefficients is due equally to different motions and different images, the variation of the amplitude

components is largely due to changes in the image, and the variations of the phase-shift components are largely due to changes in the transformation. This indicates that the amplitude components are selective for the image form and invariant to the image motion and that the transformation components are selective for the image motion and invariant to the image form.

Acknowledgments

We thank members of the Redwood Center, in particular Kilian Koepsell, Jack Culpepper, Ivana Tosic, and Jascha Sohl-Dickstein, for useful input. This work was supported by grants from the National Science Foundation (IIS-0705939) and the National Geospatial Intelligence Agency (HM1582-08-1-0007).

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America*, *2*(2), 284–299.

Albrecht, D. G., & Geisler, W. S. (1991). Motion selectivity and the contrast-response function of simple cells in the visual cortex. *Visual Neuroscience*, *7*(6), 531–546.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Barrow, H. G., & Tenenbaum, J. M. (1978). *Recovering intrinsic scene characteristics from images*. Menlo Park, CA: Artificial Intelligence Center, SRI International.

Bell, A. J., & Sejnowski, T. J. (1997). The independent components of natural images are edge filters. *Vision Research*, *37*, 3327–3338.

Berkes, P., Turner, R., & Sahani, M. (2009). A structured model of video reproduces primary visual cortical organisation. *PLoS Computational Biology*, *5*(9), e1000495.

Berkes, P., & Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision*, *5*(6), 579–602.

Born, R. T., & Bradley, D. C. (2005). Structure and function of visual area MT. *Annual Review of Neuroscience*, *28*, 157–189.

Burak, Y., Rokni, U., Meister, M., & Sompolinsky, H. (2010). Bayesian model of dynamic image stabilization in the visual system. *Proceedings of the National Academy of Sciences, USA*, *107*(45), 19525–19530.

Cadieu, C. F., & Koepsell, K. (2010a). *Modeling image structure with factorized phase-coupled Boltzmann machines*. arXiv:1011.4058v1.

Cadieu, C. F., & Koepsell, K. (2010b). Phase coupling estimation from multivariate phase statistics. *Neural Computation*, *22*(12), 3107–3126.

Cadieu, C. F., & Olshausen, B. A. (2009). Learning transformational invariants from natural movies. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems, 21* (pp. 209–216). Cambridge, MA: MIT Press.

Chen, S. S., & Gopinath, R. A. (2000). Gaussianization. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems*, 12. Cambridge, MA: MIT Press.

Culpepper, B. J., & Olshausen, B. A. (2009). Learning transport operators for image manifolds. In Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems, 22*. Cambridge, MA: MIT Press.

DeAngelis, G. C., Ghose, G. M., Ohzawa, I., & Freeman, R. D. (1999). Functional micro-organization of primary visual cortex: Receptive field analysis of nearby neurons. *Journal of Neuroscience*, *19*(10), 4046–4064.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341.

Einhauser, W., Kayser, C., Konig, P., & Kording, K. P. (2002). Learning the invariance properties of complex cells from their responses to natural stimuli. *European Journal of Neuroscience*, *15*(3), 475–486.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation*, *6*(4), 559–601.

Fleet, D. J., & Jepson, A. D. (1990). Computation of component image velocity from local phase information. *International Journal of Computer Vision*, *5*(1), 77–104.

Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, *3*(2), 194–200.

Gibson, J. J. (1983). *The senses considered as perceptual systems*. Westport, CT: Greenwood Press.

Grimes, D. B., & Rao, R.P.N. (2005). Bilinear sparse coding for invariant vision. *Neural Computation*, *17*(1), 47–73.

Heeger, D. J. (1991). Nonlinear model of neural responses in cat visual cortex. In M. Landy & J. A. Movshon (Eds.), *Computational models of visual processing* (pp. 119–133). Cambridge, MA: MIT Press.

Horn, B., & Schunck, B. (1981). Determining optical flow. *Artificial Intelligence*, *17*, 185–203.

Hoyer, P. O., & Hyvärinen, A. (2002). A multi-layer sparse coding network learns contour coding from natural images. *Vision Research*, *42*(12), 1593–1605.

Hubel, D., & Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, *195*(1), 215–243.

Hyvärinen, A., Gutmann, M., & Hoyer, P. O. (2005). Statistical model of natural stimuli predicts edge-like pooling of spatial frequency channels in V2. *BMC Neuroscience*, *6*(1), 12.

Hyvärinen, A., & Hoyer, P. O. (2000). Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, *12*(7), 1705–1720.

Hyvärinen, A., Hoyer, P. O., & Inki, M. (2001). Topographic independent component analysis. *Neural Computation*, *13*(7), 1527–1558.

Hyvärinen, A., Hurri, J., & Väyrynen, J. (2003). Bubbles: A unifying framework for low-level statistical properties of natural image sequences. *Journal of the Optical Society of America A*, *20*(7), 1237–1252.

Jhuang, H., Serre, T., Wolf, L., & Poggio, T. (2007). A biologically inspired system for action recognition. In *11th International Conference on Computer Vision (ICCV)* (pp. 1–8). Piscataway, NJ: IEEE.

Karklin, Y., & Lewicki, M. S. (2003a). Learning higher-order structures in natural images. *Network: Computation in Neural Systems*, *14*, 483–499.

Karklin, Y., & Lewicki, M. S. (2003b). A model for learning variance components of natural images. In S. Becker, S. Thrün, & K. Obermayer (Eds.), *Advances in neural information processing systems, 15* (pp. 1367–1374). Cambridge, MA: MIT Press.

Karklin, Y., & Lewicki, M. S. (2005). A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Computation*, *17*(2), 397–423.

Karklin, Y., & Lewicki, M. S. (2006). Is early vision optimized for extracting higher-order dependencies? In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems, 18*. Cambridge, MA: MIT Press.

Karklin, Y., & Lewicki, M. S. (2008). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, *457*(7225), 83–86.

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, *71*(3), 856–867.

Koenderink, J. J., & Van Doorn, A. J. (1997). The generic bilinear calibration-estimation problem. *International Journal of Computer Vision*, *23*(3), 217–234.

Koster, U., & Hyvärinen, A. (2007). A two-layer ICA-like model estimated by score matching. *Lecture Notes in Computer Science*, *4669*, 798–807.

LeCun, Y., Huang, F. J., & Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 97–104). Washington, DC: IEEE Computer Society.

Lee, H., Ekanadham, C., & Ng, A. (2007). Sparse deep belief net model for visual area V2. In J. C. Platt, D. Köller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems, 20*. Cambridge, MA: MIT Press.

Lyu, S., & Simoncelli, E. P. (2009). Nonlinear extraction of independent components of natural images using radial gaussianization. *Neural Computation*, *21*(6), 1485–1519.

Magarey, J., & Kingsbury, N. G. (1998). Motion estimation using a complex-valued wavelet transform. *IEEE Transactions on Signal Processing*, *46*(4), 1069–1084.

Memisevic, R., & Hinton, G. (2007). Unsupervised learning of image transformations. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 1–8). Washington, DC: IEEE Computer Society.

Nowlan, S. J., & Sejnowski, T. J. (1995). A selection model for motion processing in area MT of primates. *Journal of Neuroscience*, *15*(2), 1195–1214.

Olshausen, B. A. (2002). *Probabilistic models of perception and brain function*. Cambridge, MA: MIT Press.

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, *13*(11), 4700–4719.

Olshausen, B. A., Cadieu, C. F., & Warland, D. K. (2009). Learning real and complex overcomplete representations from the statistics of natural images. In *Proceedings of SPIE* (Vol. 7446, pp. 74460S–74460S-1D). Bellingham, WA: SPIE.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, *37*(23), 3311–3325.

Orban, G. A. (2008). Higher order visual processing in macaque extrastriate cortex. *Physiological Reviews*, *88*(1), 59–89.

Pollen, D. A., & Ronner, S. F. (1981). Phase relationships between adjacent simple cells in the visual cortex. *Science*, *212*(4501), 1409–1411.

Ranzato, M., & Hinton, G. (2010). Modeling pixel means and covariances using factorized third-order Boltzmann machines. In *Proc. of Computer Vision and Pattern Recognition Conference (CVPR 2010)*.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*, 1019–1025.

Rolls, E. T., & Stringer, S. M. (2007). Invariant global motion recognition in the dorsal visual system: A unifying theory. *Neural Computation*, *19*(1), 139–169.

Rust, N. C., Mante, V., Simoncelli, E. P., & Movshon, J. A. (2006). How MT cells analyze the motion of visual patterns. *Nature Neuroscience*, *9*(11), 1421–1431.

Schwartz, O., Sejnowski, T. J., & Dayan, P. (2006). Soft mixer assignment in a hierarchical generative model of natural scene statistics. *Neural Computation*, *18*, 2680–2718.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience*, *4*(8), 819–825.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*, 411–426.

Shan, H., Zhang, L., & Cottrell, G. W. (2007). Recursive ICA. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems, 19* (pp. 1273–1280). Cambridge, MA: MIT Press.

Simoncelli, E. P. (1993). *Distributed analysis and representation of visual motion*. Unpublished doctoral dissertation, MIT.

Simoncelli, E. P. (1997). Statistical models for images: Compression, restoration and synthesis. In *Conference Record of the 31st Asilomar Conference on Signals, Systems and Computers* (Vol. 1, pp. 673–678). Piscataway, NJ: IEEE.

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research*, *38*(5), 743–761.

Sinz, F., & Bethge, M. (2008). The conjoint effect of divisive normalization and orientation selectivity on redundancy reduction in natural images. In *Frontiers in Computational Neuroscience. Conference Abstract: Bernstein Symposium*. doi:10.3389/conf.neuro.10.2008.01.116.

Suchow, J. W., & Alvarez, G. A. (2011). Motion silences awareness of visual change. *Current Biology*, *21*, 140–143.

Tenenbaum, J., & Freeman, W. (2000). Separating style and content with bilinear models. *Neural Computation*, *12*(6), 1247–1283.

Tomasi, C., & Kanade, T. (1992). Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, *9*(2), 137–154.

Tsunoda, K., Yamane, Y., Nishizaki, M., & Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience*, *4*(8), 832–838.

Ullman, S. (2000). *High-level vision: Object recognition and visual cognition*. Cambridge, MA: MIT Press.

Van Hateren, J. H., & Van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society B: Biological Sciences*, *265*(1394), 359–366.

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, *51*(2), 167–194.

Wegmann, B., & Zetzsche, C. (1990). Statistical dependence between orientation filter outputs used in a human-vision-based image code. In *Proceedings of SPIE* (Vol. 1360, p. 909). Bellingham, WA: SPIE.

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, *14*(4), 715–770.

Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z., & Connor, C. E. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature Neuroscience*, *11*(11), 1352–1360.

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

Zhang, K., Sereno, M. I., & Sereno, M. E. (1993). Emergence of position-independent detectors of sense of rotation and dilation with Hebbian learning: An analysis. *Neural Computation*, *5*(4), 597–612.

Received July 6, 2011; accepted October 5, 2011.