LETTERS

Emergence of complex cell properties by learning to generalize in natural scenes

Yan Karklin[1]† & Michael S. Lewicki[1]†

A fundamental function of the visual system is to encode the building blocks of natural scenes—edges, textures and shapes—that subserve visual tasks such as object recognition and scene understanding. Essential to this process is the formation of abstract representations that generalize from specific instances of visual input. A common view holds that neurons in the early visual system signal conjunctions of image features[1,2], but how these produce invariant representations is poorly understood. Here we propose that to generalize over similar images, higher-level visual neurons encode statistical variations that characterize local image regions. We present a model in which neural activity encodes the probability distribution most consistent with a given image. Trained on natural images, the model generalizes by learning a compact set of dictionary elements for image distributions typically encountered in natural scenes. Model neurons show a diverse range of properties observed in cortical cells. These results provide a new functional explanation for nonlinear effects in complex cells[3–6] and offer insight into coding strategies in primary visual cortex (V1) and higher visual areas.

As we scan across a complex natural scene, fixations at multiple locations (for example, on the trunk of a tree or along its edge) produce a coherent percept of the underlying structure (the bark texture or the contour of the edge), even though individual images collected at the retina are inherently highly variable. Figure 1 illustrates the problem our brain solves so effortlessly: perceptually distinct image regions produce response patterns that are highly overlapping and cannot be easily distinguished using low-level, linear representations. What sort of computations are required to achieve generalization across natural stimuli?

Early visual neurons are typically described as linear feature detectors[1,2]. Models developed around this idea can accurately capture the behaviour of neurons from the retina[7] to simple cells in the cortex[8] but, as the examples in Fig. 1 illustrate, neither individual features nor linear transformations can reliably discriminate images of one structure from another. More abstract features are presumably computed in later stages of the visual system, but our knowledge of processing by these neurons is limited. In V1, complex cells respond to an edge over a range of positions[1], but classical models of these cells[9,10] fail to explain a number of nonlinear effects, such as surround suppression and cross-orientation inhibition[3–5]. More importantly, there is no functional explanation for the role of these behaviours in the perception of natural scenes. In higher visual areas such as V2 and V4, neurons are more invariant to image properties such as position and scale[11–13] and might be encoding shape or texture[12,14,15]. For these neurons to generalize effectively, the neural circuitry must generate a representation that is similar across the wide distribution of images of a given type (for example, a texture or contour) yet distinct across the much larger distribution of all other images.

Previous theoretical work has shown that neurons in the primary visual cortex form an efficient code adapted to the statistics of natural images[16,17], but this says nothing about how neurons generalize across

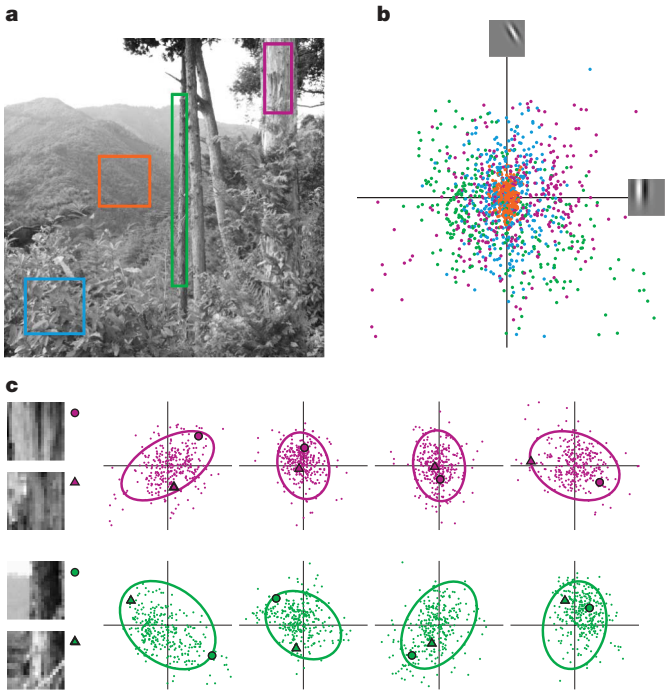

Figure 1 | Statistical patterns distinguish local regions of natural scenes. **a**, A natural scene with four distinct regions outlined (image courtesy of E. Doi). **b**, The scatter plot shows the joint output of a pair of linear feature detectors (oriented Gabor filters) for 20 × 20-image patches sampled from the four regions. The outputs from different regions are highly overlapping, indicating that linear features provide no means to distinguish between the regions. **c**, Each column shows the joint output of a different pair of linear feature detectors sampled from the regions containing the tree bark or the tree edge (the first column corresponds to features in **b**). The correlations in each panel can be described by a Gaussian distribution and its covariance (ellipses). The differences in the distributions between the rows reveal characteristic patterns in correlations, which become even more prominent as projections onto more features are considered. These patterns can be used to generalize within regions while still distinguishing among them. As an example, we highlight two patches in each region, shown by the circle and triangle in each panel. Although the pairs of images are visibly quite different, each image is consistent with the distribution of the local image region. By contrasting the distributions across multiple dimensions, it is possible to infer image type for single patches, even if the patches have similar projections along some image features.

[1]Computer Science Department & Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA 15213, USA. †Present address: Center for Neural Science, New York University, New York, New York, USA (Y.K.); Electrical Engineering and Computer Science Department, Case Western University, Cleveland, Ohio, USA and Wissenschaftskolleg (Institute for Advanced Study) zu Berlin, Germany (M.S.L.).

the intrinsic variability of scene elements. Here we extend the efficient coding approach and propose that an important aspect of visual computation is to represent the myriad statistical distributions that characterize local image regions. Rather than coding the pixel intensities of a patch of texture or edge, neurons in later stages encode the image distribution (that is, the range and pattern of variability of pixel intensities or image features) that is most consistent with the input image. This allows the neural representation to generalize across individual fixations and convey more abstract properties of the image. We demonstrate that a model designed around this computational goal and optimized for natural scenes explains nonlinear properties of complex cells and neurons in higher visual areas, thereby providing a new functional interpretation for these effects.

Fundamentally, generalization is the identification of common characteristics of a class from specific instances. The goal of the proposed model is to learn the statistical distributions that characterize local image regions, such as those in Fig. 1, and identify them from individual image patches. What statistical regularities are relevant for this task? As the examples in Fig. 1 suggest, the distributions of perceptually similar images show consistent patterns in the degree of variation along some dimensions, as well as in the strength of correlations (and anti-correlations) among different feature dimensions. Although these patterns appear subtle when projected onto two dimensions, as in the examples, the full multivariate distribution, consisting of hundreds of dimensions, produces prominent statistical signatures that can be exploited by the visual system.

To determine how the model generalizes, we must specify how it represents distributions of local image regions. A simple way to summarize the patterns of correlations for a given type of image is the covariance matrix of the data. A neural code for this structure could be defined by enumerating the set of observed covariances and assigning one neuron to each pattern, but this approach presents two problems. First, local image classes are not known a priori. Second, given the limited number of neurons in the visual system, it is not feasible to represent all possible image types, let alone the combinatorial number of possible image boundaries. Instead, we propose a distributed code in which the graded activity of the neural population acts to describe a continuum of potential covariance patterns.

This distribution coding model is illustrated schematically in Fig. 2. The model represents the correlations present in local image regions with a multivariate Gaussian distribution that has a fixed mean of zero and a covariance that is a function of the neural activity (see Methods). This simple statistical description affords both the flexibility to capture a continuum of natural image distributions and mathematical simplicity for tractable parameter estimation. The model uses two sets of parameters to describe correlations in image distributions. First, the vectors $\mathbf{b}_k$ (arrows within circles) specify image features along which the encoded distribution can be expanded or contracted relative to the canonical distribution (black circle). These vectors are shared by all neurons in the model (represented by the four grey circles, each of which contains the same set of arrows). Because these vectors do not necessarily line up with the axes of the input dimensions, changes in variation along a vector can correspond to changes in the correlational pattern in many dimensions at once. Neurons in the model (y_j) describe changes along these directions using weights w_{jk}: each has a different set of weights, corresponding to an expansion or contraction along feature $\mathbf{b}_k$. A positive weight (red) means that the neuron responds to a wider range of stimuli along that direction, a negative weight (blue) means it responds to a narrower range, and a weight close to zero (grey) indicates that the neuron is neutral to this direction. The combined activation of all neurons specifies the final shape of the encoded distribution (ellipses). Given a single fixation—one input image—the model computes the neural representation (that is, the image distribution) that provides the most probable explanation of the input. The model is able to generalize over different image regions if the inferred representation is similar across a region (for example, for the pairs of patches in Fig. 2).

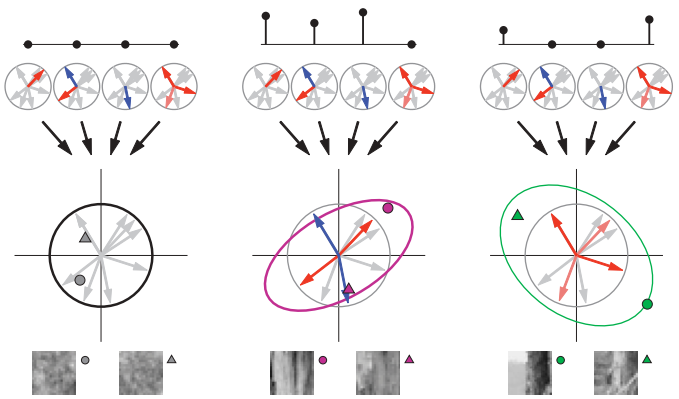

Figure 2 | Distribution coding model. Rather than encoding the precise pixel values of an input image (bottom), the proposed model infers for each image the most likely distribution (ellipses) containing it. Activation patterns for model neurons are shown at the top of each column. Absence of activity corresponds to the lack of image structure (left panel)—that is, a canonical distribution that reflects the statistics over all natural images (black circle). Increased neural activity represents deviations from this canonical distribution and captures statistical patterns in local image regions (middle and right panels, patches and symbols as in Fig. 1). In each panel, the activation pattern is the same for both inputs. See text for further details.

By adapting model parameters $\mathbf{b}_k$ and w_{jk} to the data, we are able to find the most efficient way to use a limited number of neurons to describe the wide range of distributions observed in natural images. It should be noted that, although our goal is to derive a stable representation of all patches within a local region, no assumptions about locality are made (encoding is done independently for each image patch). It is the task of the model to learn a compact representation of all patches and to automatically discover which share the statistical properties of a particular type.

If, as hypothesized, neurons in the visual cortex encode patterns in correlations in local regions and are adapted specifically to the statistics of natural scenes, we expect the representations learned by the model to reflect properties of visual neurons. To this end, we trained the model on patches sampled from a large set of natural images and examined the resulting parameters as well as the response properties of model neurons to natural images.

The vectors $\mathbf{b}_k$ encode the directions of common expansion or contraction in the shape of the image distribution. Drawn as image patches, each is an oriented and localized edge-like feature. The full set tiles the spatial extent of the image patch (Fig. 3a) and spans the range of orientations and spatial frequencies of natural images (not shown). These oriented, band-pass image features are consistent with the optimal images for exciting simple cells in the primary visual cortex[18,19]. Similar representations have been derived previously using linear statistical models that maximize the efficiency of the image codes[16,17]. In the model proposed here, however, these features are not used explicitly to reconstruct the original image, but instead function to modify the encoded distributions (arrows in Fig. 2). Thus, whereas the traditional interpretation of early sensory codes is that they are adapted for faithful reconstruction of the stimulus, our results suggest an additional interpretation: they convey variations in image distributions and allow downstream visual areas to form more abstract representations.

The second set of parameters, the weights w_{jk}, describes the role of each neuron in shaping the encoded image distribution. A set of learned weights for a typical model neuron is shown in Fig. 3b. This neuron exerts the strongest effect on features in the top left of the image patch, increasing the variability (that is, activation) of those oriented at its 'preferred' orientation of 45°, decreasing the variability of those at the orthogonal orientation, as well as those at the preferred orientation but at an offset location. Rather than

responding to a few excitatory or suppressive image features, the neuron integrates a large number to describe a pattern of variability underlying a particular image distribution. Although the functional significance of these subunits is to modify the statistical structure of the encoded distribution, they also reflect stimulus features to which this model neuron is most sensitive. It should be noted that a model neuron is activated by all images from this distribution, rather than signalling the presence of one best stimulus. Conversely, stimuli that lie in parts of image space assigned low probability by the neuron inhibit its activity.

To compare the behaviour of the model neuron to that of cells in the visual cortex, we tested its response to stimuli used in classical physiological experiments (sinusoidal gratings). Model parameters were fixed after training on natural images, and neural response computed on a set of patterns centred in the visual area that evoked maximal response. This particular model neuron showed a variety of properties observed in complex cells in V1 and cells in V2, including phase invariance, orientation tuning and complex suppressive effects (Fig. 3c). A large subset of the population exhibits similar properties, whereas others encode more complex patterns that have been observed in higher visual areas V2 and V4 (a detailed analysis of the population and similarities to other experimental data are provided in the Supplementary Information). We emphasize that these results, as well as image features described in Fig. 3a, were obtained with no assumptions about the image structure encoded by visual neurons and without fitting a model to data from physiological experiments. Specifically, we did not restrict the encoded image features to be localized and oriented, nor did we prescribe in advance how the subunits are to be combined in the pattern represented by each neuron.

Finally, we looked at the way in which the model uses the population of neurons to represent images. If the model is able to generalize across the wide variability present in natural images, then image patches that are widely scattered in the original space of linear features should be tightly clustered in the space of the model's representation. This can be illustrated by projecting into two dimensions (as was done with image space in Fig. 1) the model representation of a collection of images (Fig. 4). As hypothesized, by encoding image distributions rather than the precise feature content of each image, model neurons are able to encode perceptually similar images with similar representations and to separate distinct image types.

One limitation of the statistical framework used here is that it does not furnish an explicit feed-forward algorithm for neural encoding. Nevertheless, it is possible to approximate inference in the model by a sequential feed-forward computation: a neuron integrates the squared responses of a large number of image features $\mathbf{b}_k$ and correlates the pattern against its weights w_{jk} (see Supplementary Information for details). This computation can be viewed as a generalization of the standard model of complex cells, in which each complex cell takes as input the squared output of two simple cells[9,10,20,21]. In contrast, model neurons can receive many inputs, and the linear features themselves are learned. We find that the optimal number of input features varies greatly, and the features are integrated in a variety of ways. These predictions are consistent with recent analyses of functional subfields in V1 complex cells[6,22]. In addition, some model neurons integrate more complex spatial patterns (see Supplementary Information), which predicts a neural response to a richer variety of images than has been tested

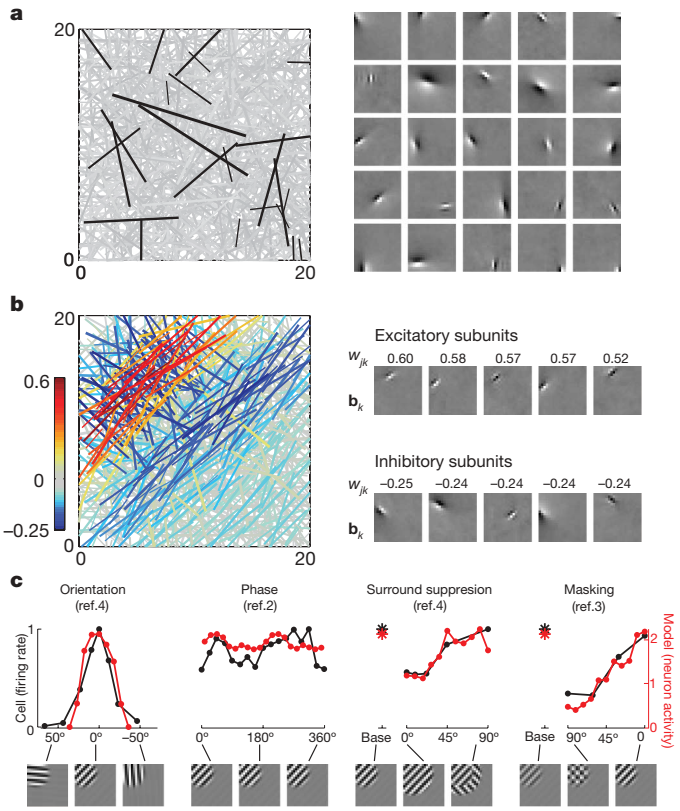

Figure 3 | Model neurons exhibit properties of cortical visual neurons. **a**, When adapted to natural images, the vectors $\mathbf{b}_k$ are oriented, localized in space, and span the spatial extent of the 20 × 20-pixel image patch. Each line reflects the orientation, spatial position within the image patch, and scale (length of line) of one of the image features. Twenty-five representative features (from a total of 1,000) are drawn in black, and shown in image form on the right. **b**, Weights of one typical model neuron to the features $\mathbf{b}_k$. As in **a**, each feature is represented by a line, and the colour of the line indicates the sign and magnitude of the weight to the feature (see colour bar). Positive weights indicate increased variability in the corresponding feature; negative weights indicate decreased variability; features to which the neuron is insensitive are shaded grey. Image features ($\mathbf{b}_k$) corresponding to the five most positive and the five most negative weights are shown in the right panel; the corresponding weights are above each feature. These act as excitatory and inhibitory subunits for this neuron. **c**, When presented with sinusoidal gratings, this model neuron replicates common aspects of the neural response in complex cells in cortical area V1. It is highly tuned to the grating's orientation, but insensitive to its phase. Adding a grating into the surrounding region suppresses the response (third plot, 0°) relative to baseline response to a single grating (asterisk), but this suppression is tuned to the orientation of the surround and is weakest when the surround is orthogonal to the preferred orientation (90°). Masking with a superimposed orthogonal grating suppresses the response (fourth plot, 90°), but this suppression is also orientation-dependent. All model neuron responses are plotted on the same scale (red axis); cell firing rates in each plot were normalized to a maximum value of 1; preferred orientation was shifted to 0° for the model neuron and the cell in all plots.

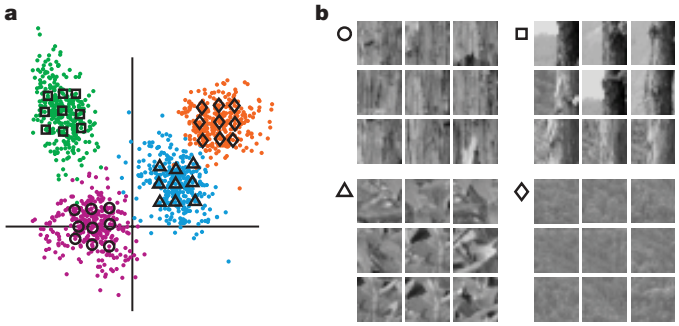

Figure 4 | Generalization across natural variability. **a**, In contrast to linear projections (compare to Fig. 1b), a two-dimensional projection of the model's representation (the activity of 150 model neurons) reveals well-separated clusters. **b**, Each 3 × 3-image group corresponds to the array of symbols in **a**. Despite the variability in the appearance of edges and textures, the model's representation of natural images generalizes within each region while still distinguishing among them.

physiologically. Experiments that incorporate such stimuli will provide an important validation of the proposed model.

The nonlinear effects shown by neurons in the model (Fig. 3c) have been previously incorporated into models of complex cells[5,8,20,21]. Much of this work has focused on fitting mathematical models to neural data[5,8,20,23] and does not provide a functional explanation of the observed neural properties. Other models have been motivated by specific computational goals, such as statistical independence[24,25], stability of representation over time[26,27], or position or scale invariance[28]. However, these models do not explicitly address the problem of generalization, which here is performed by inferring the statistical distribution that is most likely to explain the input image. An important advantage of our approach is that, rather than assuming invariance (or sensitivity) to limited stimulus parameters such as position or orientation, the model learns a much more general set of features that are determined by the statistical structures in natural images. If higher-level visual neurons are generalizing according to these statistics, they should have invariance along specific stimulus dimensions, and their responses to natural images should reflect common statistical structure in local image regions. Thus, the model provides a quantitative way to explore neural responses to complex stimuli characterized by their statistical regularities.

METHODS SUMMARY

The model describes individual image patches $\mathbf{x}$ with multivariate Gaussian probability distributions:

$$P(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}) \quad (1)$$

with mean $\boldsymbol{\mu} = 0$ and with covariance a function of the neural encoding vector $\mathbf{C} = f(\mathbf{y})$. The logarithm of the covariance matrix is given by the combination of outer products of feature vectors $\mathbf{b}_k$, weighted by neural activities y_j through weights w_{jk}:

$$\log \mathbf{C} = \sum_{jk} y_j w_{jk} \mathbf{b}_k \mathbf{b}_k^T \quad (2)$$

Because a different covariance can be inferred for each image, the distribution over the entire ensemble of images is highly non-Gaussian. (This model is a generalized version of the hierarchical model described previously[29], which captured patterns among the variances, but not the correlations, of linear features.)

We trained the model on a large set of 20×20 image patches, sampled randomly from greyscale photographs of outdoor scenes[19]. The number of neurons was set to 150 and the number of linear features $\mathbf{b}_k$ to 1,000. The 'response' of model neurons was computed as the most probable neural representation given the input image by maximizing the posterior probability $P(\mathbf{y}|\mathbf{x}, \{\mathbf{b}_k, w_{jk}\})$. Model parameters were initialized to small random values and optimized by maximizing the likelihood of the data under the model $P(\mathbf{x}|\{\mathbf{b}_k, w_{jk}\})$ using standard gradient ascent.

For the 'physiological' analysis of Fig. 3c, we first identified the location, orientation, and spatial extent and frequency of a windowed sinusoidal grating that best activated the model neuron (one that yielded the most positive value of y_j). We then varied each tested parameter and computed the model's representation of the stimulus (the vector of responses of model neurons).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 4 May; accepted 26 September 2008.
Published online 19 November 2008.

1. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)* **160,** 106–154 (1962).
2. Movshon, J. A., Thompson, I. D. & Tolhurst, D. J. Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *J. Physiol. (Lond.)* **283,** 53–77 (1978).
3. Bonds, A. B. Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. *Vis. Neurosci.* **2,** 41–55 (1989).
4. Jones, H. E., Wang, W. & Sillito, A. M. Spatial organization and magnitude of orientation contrast interactions in primate V1. *J. Neurophysiol.* **88,** 2796–2808 (2002).
5. Cavanaugh, J. R., Bair, W. & Movshon, J. A. Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *J. Neurophysiol.* **88,** 2530–2546 (2002).
6. Chen, X., Han, F., Poo, M. & Dan, Y. Excitatory and suppressive receptive field subunits in awake monkey primary visual cortex (V1). *Proc. Natl Acad. Sci. USA* **104,** 19120–19125 (2007).
7. Chichilnisky, E. J. A simple white noise analysis of neuronal light responses. *Network: Comp. Neural Syst.* **12,** 199–213 (2001).
8. Carandini, M., Heeger, D. J. & Movshon, J. A. Linearity and normalization in simple cells of the macaque primary visual cortex. *J. Neurosci.* **17,** 8621–8644 (1997).
9. Movshon, J. A., Thompson, I. D. & Tolhurst, D. J. Receptive field organization of complex cells in the cat's striate cortex. *J. Physiol. (Lond.)* **283,** 79–99 (1978).
10. Adelson, E. H. & Bergen, J. R. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* **2,** 284–299 (1985).
11. Kobatake, E. & Tanaka, K. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.* **71,** 856–867 (1994).
12. Gallant, J. L., Connor, C. E., Rakshit, S., Lewis, J. W. & Van Essen, D. C. Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *J. Neurophysiol.* **76,** 2718–2739 (1996).
13. Connor, C. E., Brincat, S. L. & Pasupathy, A. Transformation of shape information in the ventral pathway. *Curr. Opin. Neurobiol.* **17,** 140–147 (2007).
14. Hegdé, J. & Van Essen, D. C. Selectivity for complex shapes in primate visual area V2. *J. Neurosci.* **20,** RC61:1–6 (2000).
15. Pasupathy, A. & Connor, C. E. Shape representation in area V4: position-specific tuning for boundary conformation. *J. Neurophysiol.* **86,** 2505–2519 (2001).
16. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381,** 607–609 (1996).
17. Bell, A. J. & Sejnowski, T. J. The ''independent components'' of natural scenes are edge filters. *Vision Res.* **37,** 3327–3338 (1997).
18. Jones, J. P. & Palmer, L. A. The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **58,** 1187–1211 (1987).
19. van Hateren, J. H. & van der Schaaf, A. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. Lond. B* **265,** 359–366 (1998).
20. Heeger, D. J. Normalization of cell responses in cat striate cortex. *Vis. Neurosci.* **9,** 181–197 (1992).
21. Heeger, D. J., Simoncelli, E. P. & Movshon, J. A. Computational models of cortical visual processing. *Proc. Natl Acad. Sci. USA* **93,** 623–627 (1996).
22. Rust, N. C., Schwartz, O., Movshon, J. A. & Simoncelli, E. P. Spatiotemporal elements of macaque V1 receptive fields. *Neuron* **46,** 945–956 (2005).
23. Cadieu, C. *et al.* A model of V4 shape selectivity and invariance. *J. Neurophysiol.* **98,** 1733–1750 (2007).
24. Schwartz, O. & Simoncelli, E. P. Natural signal statistics and sensory gain control. *Nature Neurosci.* **4,** 819–825 (2001).
25. Hyvärinen, A. & Hoyer, P. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Res.* **41,** 2413–2423 (2001).
26. Berkes, P. & Wiskott, L. Slow feature analysis yields a rich repertoire of complex cell properties. *J. Vis.* **5,** 579–602 (2005).
27. Hurri, J. & Hyvärinen, A. Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Comput.* **15,** 663–691 (2003).
28. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nature Neurosci.* **2,** 1019–1025 (1999).
29. Karklin, Y. & Lewicki, M. S. A hierarchical Bayesian model for learning non-linear statistical regularities in non-stationary natural signals. *Neural Comput.* **17,** 397–423 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the Department of Energy through the Computational Science Graduate Fellowship (to Y.K.), the National Science Foundation Grant under grant numbers 0413152 and 0705677 (to M.S.L.) and the Office of Naval Research under the Multidisciplinary University Research Initiative N000140710747.

Author Contributions Y.K. and M.S.L. developed the model, analysed the results and wrote the paper; Y.K. ran the simulations.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to Y.K. (yan.karklin@nyu.edu) or M.S.L. (michael.lewicki@case.edu).

METHODS

Data. We used 110 greyscale images of outdoor scenes as training data[19]. Pixel intensities were log-transformed (corresponding roughly to the transformation at the retinal cone cells[30]), and the images were low-pass filtered to remove corner frequency sampling artefacts. We randomly extracted overlapping 20×20-image patches from the entire data set. The mean luminance value was subtracted from each patch (which sped up model training but had no significant influence on the results). We 'whitened' all image patches to remove global correlations and to normalize the variance; this allowed the model to encode only the deviations of each image distribution from the global statistics (the canonical distribution). For visualization of image features, the results were projected back into the original image space. All stimuli in the physiological analysis of Fig. 3c were preprocessed in the same way as the natural images used in training.

Model parameter estimation. We estimated the optimal model parameters $\theta = \{\mathbf{b}_k, w_{jk}\}$ by maximizing the likelihood of the data under the model

$$P(\mathbf{x}|\theta) = \int P(\mathbf{x}|\mathbf{y}, \theta)P(\mathbf{y})d\mathbf{y} \quad (3)$$

The conditional distribution $P(\mathbf{x}|\mathbf{y}, \theta)$ is a multivariate Gaussian that captures correlations in local image regions (equation (1)). Neural activities were assumed to be sparse[31] and independent, and were modelled with a Laplacian (symmetric exponential) distribution, $P(\mathbf{y}) = \Pi P(y_j) \propto \Pi e^{-|y_j|}$. The integral over all possible neural states in equation (3) is intractable and was replaced by a single evaluation at the maximum a posteriori value $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}, \theta)$. Although this approximation ignores the volume around the maximum, it is one standard approach to tackling this problem.

We assumed the training patches were sampled independently and that the likelihood for the data ensemble was a product of terms for individual images (equation (3)). In practice, we maximized the log-likelihood using gradient ascent on batches of 100 image patches. Repeated training runs produced convergence to similar parameter values.

Model responses to grating stimuli. The orientation tuning of model neurons in Fig. 3c was measured using 20×20 patches of sinusoidal gratings at different positions, orientations, spatial frequencies and phases. We first eliminated neurons that were 'unresponsive' to gratings, that is, those whose maximal response did not reach 2 standard deviations of the population response to gratings. This was necessary to discount small random activation of neurons tuned for other types of image structures. For each neuron we found the grating with the maximal response and measured modulation in response to varying orientation, phase, or the addition of masks in the receptive field or the surround. Because neural activity could be positive or negative, the full amplitude of modulation was considered as twice the maximum absolute value of the response.

A neuron was considered to be orientation-tuned if its response was modulated by more than 50% over the range of stimulus orientations, and to be phase invariant if the response varied less than 50% over phase-shifted gratings. Cross-orientation inhibition and surround suppression corresponded to greater than 25% decrease in neural response. Bandwidth of orientation tuning was computed as the width at $1/\sqrt{2}$ of the full amplitude of the response modulation.

The projection of neural activity in Fig. 4 was computed using linear discriminant analysis, a technique that finds the linear projections that best separate different classes of data. Applied to the raw pixel data or to the outputs of linear features (data shown in Fig. 1), this method failed to separate the clusters.

30. van Hateren, J. H. Processing of natural time series of intensities by the visual system of the blowfly. *Vision Res.* **37,** 3407–3416 (1997).
31. Olshausen, B. A. & Field, D. J. Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* **14,** 481–487 (2004).