A Model of V4 Shape Selectivity and Invariance

Charles Cadieu, Minjoon Kouh, Anitha Pasupathy, Charles E. Connor, Maximilian Riesenhuber and Tomaso Poggio

J Neurophysiol 98:1733-1750, 2007. First published 27 June 2007; doi:10.1152/jn.01265.2006

J Neurophysiol 98: 1733–1750, 2007.
First published June 27, 2007; doi:10.1152/jn.01265.2006.

A Model of V4 Shape Selectivity and Invariance

Charles Cadieu,[1] Minjoon Kouh,[1] Anitha Pasupathy,[2] Charles E. Connor,[3] Maximilian Riesenhuber,[4] and Tomaso Poggio[1]

[1]*Center for Biological and Computational Learning, McGovern Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts;* [2]*Department of Biological Structure, University of Washington, Seattle, Washington;* [3]*Department of Neuroscience, Johns Hopkins University, Baltimore, Maryland; and* [4]*Department of Neuroscience, Georgetown University Medical Center, Washington, DC*

Submitted 2 December 2006; accepted in final form 24 June 2007

Cadieu C, Kouh M, Pasupathy A, Connor CE, Riesenhuber M, Poggio T. A model of V4 shape selectivity and invariance. *J Neurophysiol* 98: 1733–1750, 2007. First published June 27, 2007; doi:10.1152/jn.01265.2006. Object recognition in primates is mediated by the ventral visual pathway and is classically described as a feedforward hierarchy of increasingly sophisticated representations. Neurons in macaque monkey area V4, an intermediate stage along the ventral pathway, have been shown to exhibit selectivity to complex boundary conformation and invariance to spatial translation. How could such a representation be derived from the signals in lower visual areas such as V1? We show that a quantitative model of hierarchical processing, which is part of a larger model of object recognition in the ventral pathway, provides a plausible mechanism for the translation-invariant shape representation observed in area V4. Simulated model neurons successfully reproduce V4 selectivity and invariance through a nonlinear, translation-invariant combination of locally selective subunits, suggesting that a similar transformation may occur or culminate in area V4. Specifically, this mechanism models the selectivity of individual V4 neurons to boundary conformation stimuli, exhibits the same degree of translation invariance observed in V4, and produces observed V4 population responses to bars and non-Cartesian gratings. This work provides a quantitative model of the widely described shape selectivity and invariance properties of area V4 and points toward a possible canonical mechanism operating throughout the ventral pathway.

INTRODUCTION

Visual object recognition is a computationally demanding task that is frequently performed by the primate brain. Primates are able to discriminate and recognize objects under a variety of conditions, such as changes in position, rotation, and illumination, at a level of proficiency and speed that is currently unmatched by engineered systems. How the primate brain achieves this level of proficiency has largely been unexplained, but it seems clear that the computations employed must achieve both selectivity and invariance. In light of this computational requirement, neurons in visual area V4 have been shown to exhibit responses that are both selective for complex stimuli and invariant to spatial translations (Desimone and Schein 1987; Freiwald et al. 2004; Gallant et al. 1996; Kobatake and Tanaka 1994). Moreover, visual area V4 is likely to play a critical role in object recognition: i.e., a lesion in this area results in the impairment of shape perception and attention (De Weerd et al. 1996; Gallant et al. 2000; Girard et al. 2002; Merigan and Pham 1998; Schiller 1995; Schiller and Lee 1991).

Area V4 lies in the middle of the ventral pathway, which is one of two major cortical pathways that process visual information and which has been closely linked to object recognition by a variety of experiments (for a review, see Ungerleider and Haxby 1994). Several studies have explored and described the representations at various stages along the ventral pathway (Kobatake and Tanaka 1994). These studies have shown that the responses of neurons in lower visual areas, such as primary visual cortex (V1), and higher visual areas, such as inferotemporal (IT) complex, explicitly represent features or information about visual form. Neurons in the early stages of the ventral pathway in V1 have small receptive fields and are responsive to simple features, such as edge orientation (De Valois et al. 1982; Hubel and Wiesel 1962), whereas neurons far along the pathway in IT have large receptive fields and can be selective for complex shapes like faces, hands, and specific views of other familiar objects (Gross et al. 1972; Hung et al. 2005; Logothetis et al. 1995; Tanaka et al. 1991). Neural response properties in area V4 reflect its intermediate anatomical position. V4 receptive field sizes average four to seven times those in V1 but are smaller than those in IT (Desimone and Schein 1987; Kobatake and Tanaka 1994). Many V4 neurons are sensitive to stimulus features of moderate complexity (Desimone and Schein 1987; Freiwald et al. 2004; Gallant et al. 1996; Gawne and Martin 2002; Kobatake and Tanaka 1994; Pasupathy and Connor 1999, 2001; Pollen et al. 2002).

Previously, Pasupathy and Connor (1999, 2001) provided a quantitative, phenomenological description of stimulus shape selectivity and position invariance in area V4. They demonstrated that a subpopulation of V4 neurons, screened for their high firing rates to complex stimuli, is sensitive to local modulations of boundary shape and orientation (Pasupathy and Connor 1999). The responses of these neurons can be described as basis function-like tuning for curvature, orientation, and object-relative position of boundary fragments within larger, more complex global shapes (Pasupathy and Connor 2001). This tuning is relatively invariant to local translation. At the population level, a global shape may be represented in terms of its constituent boundary fragments by multiple peaks in the population response pattern (Pasupathy and Connor 2002). Brincat and Connor showed that V4 signals for local boundary fragments may be integrated into more complex shape constructs at subsequent processing stages in posterior IT (Brincat and Connor 2004, 2006).

Physiological findings in V4 and other areas of the ventral stream have led to a commonly held belief about how object

Present address and address for reprint requests and other correspondence: C. Cadieu, Redwood Center for Theoretical Neuroscience, University of California, Berkeley, Helen Wills Neuroscience Institute, 132 Barker Hall, 3190, Berkeley, CA 94720-3190 (E-mail: cadieu@berkeley.edu)

recognition is achieved in the primate brain and specifically how selectivity and invariance could be achieved in area V4. Hubel and Wiesel first recognized selectivity and invariance by probing neurons in cat area 17 with Cartesian gratings and oriented bars. They found that some cells (classified as "simple") exhibited strong phase dependence, whereas others (classified as "complex") did not. Hubel and Wiesel proposed that the invariance of those complex cells they described could be formed by pooling together simple cells with similar selectivities but with translated receptive fields (Hubel and Wiesel 1962, 1965). Perrett and Oram (1993) proposed a similar mechanism within IT to achieve invariance to any transformation by pooling afferents tuned to transformed versions of the same stimuli. Based on these hypotheses, quantitative models of the ventral pathway have been developed (Fukushima et al. 1983; Mel 1997; Riesenhuber and Poggio 1999; Serre et al. 2005, 2007a) with the goal of explaining object recognition. The V4 model presented here is part of a model (Serre et al. 2005, 2007a) of the entire ventral pathway. Within this framework, we sought to explain the observed response characteristics of V4 neurons described in Pasupathy and Connor (2001) (selectivity for boundary fragment conformation and object-relative position and invariance to local translations) in terms of a biologically plausible, feedforward model of the ventral pathway motivated by the computational goal of object recognition.

Our V4 model shows that the response patterns of V4 neurons described in Pasupathy and Connor (2001) can be quantitatively reproduced by a translation-invariant combination of locally selective inputs. Simulated responses correspond closely to physiologically measured V4 responses of individual neurons during the presentation of stimuli that test selectivity for complex boundary conformation and invariance to local translation. The model provides a possible explanation of the transformation from lower level visual areas to the responses observed in V4. Model neurons can also predict physiological responses to stimuli that were not used to derive the model, allowing for comparison with other independent experimental results. The model neurons and their corresponding V4 neuron population may be interpreted on a geometric level as boundary conformation filters, just as V1 neurons can be considered edge or orientation filters.

METHODS

Model of V4 shape representation

The model is motivated by a theory of object recognition (Riesenhuber and Poggio 1999; Serre et al. 2005, 2007a) and its parameters that are specific to V4 incorporate neurophysiological evidence (Pasupathy and Connor 2001). These considerations motivate four major aspects of the model. First, the architecture of the model is hierarchical, reflecting the anatomical structure of the primate visual cortex (Felleman and Van Essen 1991). Second, the main computations are feedforward, as suggested by results of rapid categorization/recognition experiments, such as (Hung et al. 2005; Thorpe et al. 1996). Third, the V1-like layers of the model are composed of orientation-tuned, Gabor-filtering units that match observed physiological evidence in V1 (Serre et al. 2005). Finally, two computations are performed in alternating layers of the hierarchy, mimicking the observed, gradual build-up of shape selectivity and invariance along the ventral pathway. A software implementation of the full model of the ventral pathway is available at http://cbcl.mit.edu.

The key parts of the resulting V4 model are summarized schematically in Fig. 1. It comprises four layers: S1, C1, S2, and C2. Each layer contains either "S" units performing a selectivity operation on their afferents or "C" units performing an invariance operation on their afferents. The lower S1, C1, and S2 units of the model are analogous to neurons in the visual areas V1 and V2, which precede V4 in the feedforward hierarchy (the role of V2 and the issue of anatomical correspondence for the S2 layer are considered in DISCUSSION). A single C2 unit at the top level of the hierarchy models an individual V4 neuron's response.

Our V4 model is consistent with several other quantitative and qualitative models of V4 (e.g., Gallant et al. 1996; Li 1998; Reynolds et al. 1999; Wilson and Wilkinson 1998), where several orientation-selective afferent units are combined with nonlinear feedforward operations, often involving inhibitory elements. Such models have been successful in describing and explaining different specific phenomena, such as texture discrimination (Wilson and Wilkinson 1998), contour integration (Li 1998), or attentional effects (Reynolds et al. 1999), occurring in or around V4. Our model differs and extends these previous descriptions in a number of ways. First, in our model the role of area V4 is part of a framework that attempts to explain the entire ventral pathway at a computational and quantitative level. Second, our model not only attempts to explain experimental findings but also attempts to explain how V4 responses could be computed from the known properties of earlier stages within the ventral pathway. Third, our model involves two stages of computation to account for the selectivity of V4 neurons to complex stimuli and their invariance to visual translations.

Operations

The model has two main operations: the selectivity operation and the invariance operation. Selectivity is generated by a bell-shaped, template-matching operation on a set of inputs from the afferent units. A normalized dot product followed by a sigmoid function is used as a biologically plausible implementation of the selectivity operation. This operation can be implemented with synaptic weights and an inhibitory mechanism (Poggio and Bizzi 2004; Serre et al. 2005). The response, r, of a selectivity unit (i.e., an S2 unit) is given by

$$r = g\left(\frac{\sum_i w_i x_i}{\sqrt{\sum_i x_i^2 + k}}\right), \quad (1)$$

where x_i is the response of the ith afferent unit, w_i is the synaptic weight of the ith afferent, and the sigmoid function $g(u)$ is given by

$$g(u) = \frac{s}{1 + \exp(-\alpha(u - \beta))} \quad (2)$$

The sigmoid parameters, α and β, determine the steepness of tuning, and s represents the maximum response of the unit. A small number k (0.0001) prevents division by zero. The divisive normalization in *Eq. 1* can arise from lateral or feedforward shunting inhibitions, and it is closely related to the inhibitory elements in other models of V4 [e.g., center-surround inhibition in Wilson and Wilkinson (1998) and especially the biased-competition model formulation in Reynolds et al. (1999)]. The resulting function is selective to a conjunction of the input activity and is functionally similar to a Gaussian tuning function. While similar results could be obtained with a Gaussian tuning function, it is not clear how a Gaussian function could be implemented directly with neural circuits. Hence the preceding functional form was chosen for its biophysical plausibility and is also used in the full model of Serre et al. (2005).

The invariance operation is implemented by the maximum function. The maximum response of afferents with the same

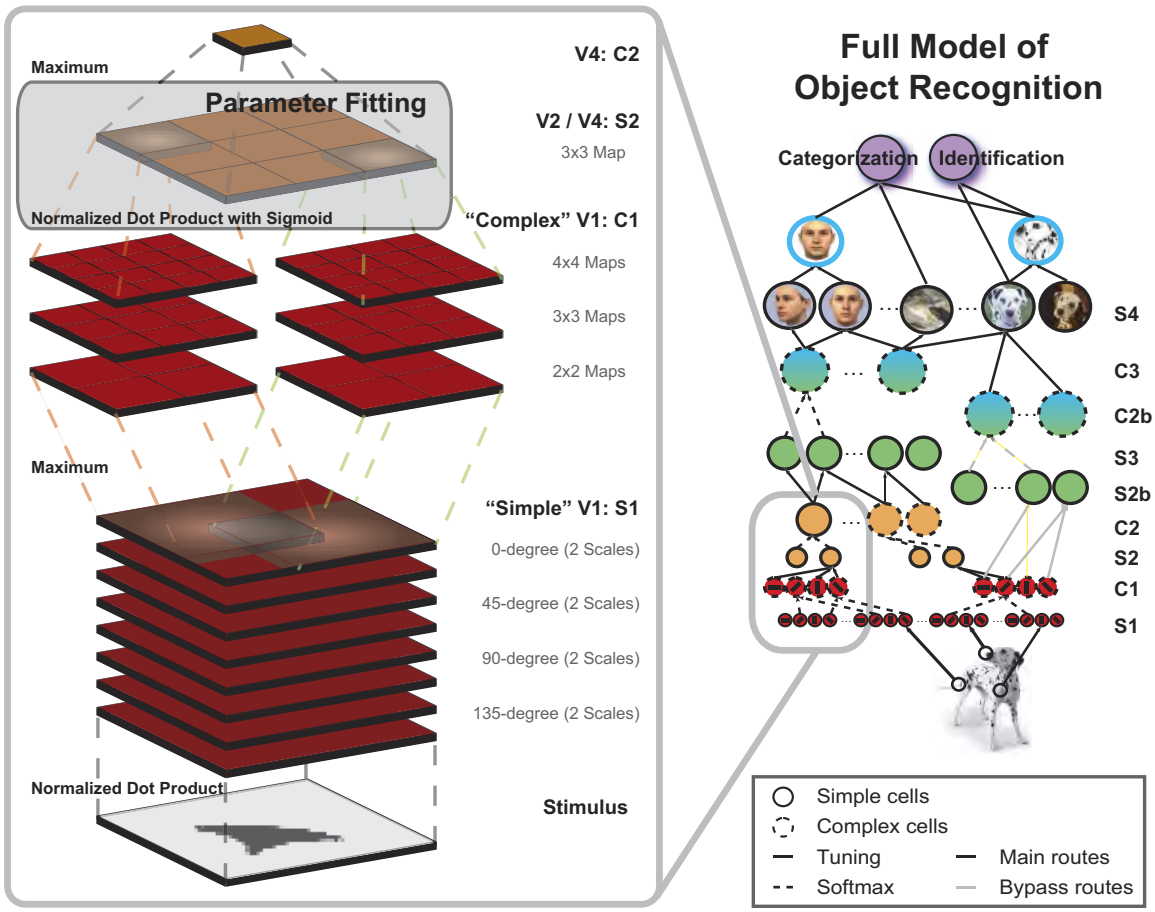

FIG. 1. Model of V4 shape representation. Our model of V4 (*left*) is part of an extensive theory of object recognition (*right*) dealing with the computations and neuronal circuits in the feedforward pathway of the ventral stream in primate visual cortex (Riesenhuber and Poggio 1999; Serre et al. 2005). The response of a C2 unit (*left top*) is used to model the responses of individual V4 neurons and is determined by the preceding layers of units, corresponding to earlier stages of the ventral pathway before area V4. The build-up of selectivity and invariance is implemented by the alternating hierarchical layers of "simple" S units, performing a selectivity operation, and "complex" C units performing an invariance operation. The designation of simple and complex follows the convention of distinguishing between the orientation-tuned, phase-dependent simple cells and the translation-invariant complex cells of V1 (Hubel and Wiesel 1962, 1968). Because V4 neurons exhibit both selectivity for complex shapes and invariance to local translation, V4 neurons are modeled with the responses of C2 units by the combination of translated copies of S2 unit afferents with identical selectivity, but shifted receptive fields, following the same construction principle as in S1 and C1 layers. The lower S1 and C1 units of the model are analogous to neurons in area V1. In the feedforward direction, the image is processed by simple V1-like S1 units that send efferent projections to complex V1-like C1 units (for clarity, only a subset of C1 units are shown). S2 units receive input from a specific combination of C1 unit afferents and are selective for a particular activation of those inputs. Finally, the C2 unit pools over shifted S2 units. The resulting C2 unit produces a high response to a specific stimulus and is invariant to the exact position of the stimulus within the receptive field (the full receptive field spans the union of the receptive fields of S1 units). For different, nonoptimal stimuli, the C2 response falls off as the afferent activity deviates from the optimal pattern. Most parameters in the model are fixed, except for the C1 and S2 connectivity (indicated by shaded rectangular region), which is varied to fit the individual neural responses. Details of the model implementation and the fitting procedures can be found in METHODS.

selectivity, but translated or scaled receptive fields, produces responses that are invariant to translation or scale. An approximate maximum operation, known as softmax, can also be performed by a normalized dot product neural circuitry similar to the selectivity operation (Serre et al. 2005; Yu et al. 2002). In the simulations described here, we used the maximum operation over afferent inputs instead of the softmax.

S1 and C1 layers

The selectivity and invariance operations are performed in alternating layers (see Fig. 1): S1, C1, S2, and C2. In the feedforward direction, the pixels of the gray level valued image are processed by S1 units that correspond to "simple" cells in V1. They have Gabor receptive field profiles with different sizes and four orientations (0, 45, 90, and 135°). The S1 filter, h, is given by the Gabor function

$$h(p_1, p_2) = \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right)\cos\left(\frac{2\pi}{\lambda}u - \phi\right)$$
$$x = p_1 \cos\theta + p_2 \sin\theta,$$
$$y = -p_1 \sin\theta + p_2 \cos\theta, \quad (3)$$

where p_1 and p_2 indicate the coordinate indices centered on the S1 unit's receptive field and range between $-\pi$ and π, θ gives the orientation of the filter, and ϕ gives the phase offset. For all S1 filters the parameters were set to: $\lambda = 2.1$, $\sigma_x = 2\pi/3$, $\sigma_y = 2\pi/1.8$, and $\phi = 0$. The responses of S1 units are the normalized dot product of the Gabor filter and the image patch within the receptive field. The sigmoid nonlinearity is not used in the S1 selectivity function. This results in a model of simple V1 neurons that is similar to that presented in Carandini et al. (1997) and Heeger (1993). S1 responses were rectified by taking their absolute value, which is equivalent to having rectified S1 units of both signs project to the same efferent units. C1 units, which correspond to complex V1 cells, perform the invariance operation (maximum function) over

S1 units with identical selectivity (i.e., orientation) but slightly shifted or scaled receptive fields. As a result of such construction, C1 units have orientation selectivity with larger receptive fields than S1 units within which we observe translation and scale invariance, similar to complex V1 cells.

Three different spatial pooling ranges over S1 units are used to create C1 units with varying receptive field sizes, as observed in V1 (Hubel and Wiesel 1962, 1968). The S1 and C1 parameters are fixed throughout all simulations and are listed in Table 1. The receptive fields of adjacent C1 units (with the same size) overlap by 50%. The parameters for S1 and C1 units have been chosen to reflect experimental findings (e.g., receptive field sizes, orientation and spatial frequency tuning, differences between spatial frequency bandwidth between simple and complex cells, etc.) about V1 neurons (De Valois et al. 1982; Schiller et al. 1976; Serre et al. 2005).

S2 and C2 layers

The same construction principle in S1 and C1 is repeated in the next two layers, S2 and C2, and the parameters are given in Table 1. S2 units perform the selectivity operation on their C1 afferents, generating selectivity for features or shapes more complex than just orientation selectivity. Within the receptive field of each S2 unit, there are C1 units with three different receptive field sizes. The C1 units with the smallest receptive field size span the S2 receptive field in a 4×4 array, whereas C1 units with larger receptive field sizes span the S2 receptive field in 3×3 or 2×2 arrays. Therefore within each S2 receptive field there are 29 $[(2 \times 2) + (3 \times 3) + (4 \times 4)]$ spatial locations, each with units at four different orientations, resulting in a total of 116 (29×4) potential C1 units that could provide an input to an S2 unit. A small subset of these 116 C1 units is connected to each S2 unit, and different combinations of C1 subunits produce a wide variety of complex shape selectivities. The selection of which C1 subunits connect to an S2 unit, their connection strengths, and the three sigmoid parameters in *Eq. 2* are the only parameters fit to a given V4 neuron's response.

The top level C2 unit, which corresponds to a V4 neuron, performs the invariance operation on the afferent projections from the S2 layer. Because V4 neurons exhibit both selectivity for complex shapes and invariance to local translation, V4 neurons are likely to combine translated copies of inputs with the same, but shifted, selectivity, just like the construction of a V1 complex cell. According to experimental studies (Desimone and Schein 1987; Gallant et al. 1996; Pasupathy and Connor 1999, 2001), V4 neurons maintain selectivity to translations of ~0.5 times the classical receptive field size. To match these experimental findings, a C2 unit receives input from a 3×3 spatial grid of S2 units with identical selectivity properties, each shifted by 0.25 times the S2 receptive field (i.e., 1 C2 unit receives inputs from 9 S2 units). As a result, the C2 unit adopts the selectivity of its afferent S2 units to a particular pattern evoked by a stimulus in C1 and is invariant to the exact position of the stimulus. The C2 parameters, controlling the receptive field size and the range of translation invariance, are fixed throughout all the simulations.

In summary, our model of V4 is composed of hierarchical layers of model units performing feedforward selectivity or invariance operations. Most of the parameters are fixed to reasonable estimates based on experimental data from areas V1 and V4. To model a particular V4 neuron, only the parameters governing the connectivity between C1 and S2 layers, as indicated by the shaded rectangular region in Fig. 1, are found according to the fitting technique described in the *Fitting model parameters* section.

The current version of the model (Serre et al. 2005) is an extension of the original formulation (Riesenhuber and Poggio 1999) in three ways: the optimal activation patterns for S2 units are more varied to account for the diverse selectivity properties measured in V4, the tuning operation for the S2 layer has a more biologically plausible form, *Eq. 1*, and the max-pooling range for the C2 layer is set to match the invariance properties of V4 neurons. These changes were natural and planned extensions of the original model. Further information can be found in (Serre et al. 2005). The full version of the model (Serre et al. 2005, 2007a) has additional layers above C2 that are comparable to the higher areas of the visual cortex like posterior and anterior inferotemporal cortex and prefrontal cortex, and complete the hierarchy for functional object recognition. The full model also sets the tuning of the S2 and S3 units with an unsupervised learning stage using thousands of natural images. These modifications do not change the results of the analysis in (Riesenhuber and Poggio 1999) of responses of neurons in IT (Cadieu et al. 2004).

Physiological data

Using our model of V4, we examined the electrophysiological responses of 109 V4 neurons previously reported in Pasupathy and Connor (2001). The stimulus set construction and the physiological methods are fully described in Pasupathy and Connor (2001). Briefly, the stimulus set was designed to be a partial factorial cross of boundary conformation values (sharp to shallow convex and concave curvature) at 45°-interval angular positions (relative to object center). The factorial cross is only partial because a complete cross is geometrically impossible without creating boundary discontinuities that would result in irregular shapes (for example, a closed contour shape cannot be generated by using concave curvatures only). Responses of individual neurons were recorded from parafoveal V4 cortex of awake, fixating monkeys (*Macaca mulatta*) using standard electrophysiological techniques. The response to each stimulus shape during a 500-ms presentation period was averaged across three to five repetitions. For the analyses presented here, each neuron's responses across the entire stimulus set were normalized to range between 0 and 1.

Fitting model parameters

For each V4 neuron, we wanted to determine parameters within the model that would produce matching responses to that neuron's selectivity and invariance profile. Although a number of parameters could

TABLE 1. *Model parameters*

SCALE	S1 RF	C1 RF	C1 SHIFT	C1 GRID	S2 RF	S2 SHIFT	S2 GRID	C2 RF
1	54,60	80	40	2×2				
2	40,45	60	30	3×3	120	30	3×3	180
3	32,36	48	24	4×4				

The S1 and C1 layers are broken down into three spatial scales (1st column). The receptive field (RF) sizes of S1 units vary across spatial scales (2nd column, measured in pixels). Within each spatial scale, CI units receive input from S1 units with 2 different RF sizes (to achieve a small degree of scale invariance) and with different spatial locations (to achieve translation and phase invariance). The resulting RF sizes of CI units are indicated in the 3rd column. Within each spatial scale, CI units form spatial grids with the center of adjacent CI receptive fields shifted by the amount indicated in the 4th column. In the S2 layer, S2 units receive input from all 3 spatial scales. The CI grid sizes for each spatial scale span the same range of space (120×120 pixels), giving S2 units an identical RF size. S2 units form a 3×3 grid with adjacent S2 units shifted by 30 pixels. The top layer C2 unit has a RF size of 180 pixels. In the model, 32 pixels correspond to ~1° of visual angle.

be adjusted to accomplish this goal, the selectivity of a C2 unit, which corresponds to a V4 neuron, is most dependent on the spatial arrangement and synaptic weights connecting C1 units to the S2 units (modifying other parameters had little effect on the level of fit, see *Increasing the parameter space*). Furthermore, the model layers before S2 were not adjusted because they are considered analogous to representations in V1 and were not the focus of this study. The invariance operation at the C2 layer was not adjusted because experimental results indicate that translation invariance over measured V4 populations is highly consistent (Desimone and Schein 1987; Gallant et al. 1996; Pasupathy and Connor 1999, 2001) and because the experimental measurements modeled here do not include sufficient stimuli at different translations. Therefore the fitting algorithm determined the parameters of the selectivity operation at the S2 layer while holding all other parameters fixed (the fitted parameters within the overall model are indicated by the shaded box in Fig. 1, *left,* labeled as "parameter fitting"). Specifically, these parameters included the subset of C1 afferents connected to an S2 unit, the connection weights to those C1 afferents, and the parameters of the sigmoid function that nonlinearly scaled the response values. For a given C2 unit, the parameters for all 3×3 afferent S2 units were identical to produce identical tuning over translation.

Because the model's hierarchy of nonlinear operations makes analytical solutions intractable, we used numerical methods to find solutions. For each C2 unit, we needed to determine the set of C1 subunits connected to the S2 units, the weights of the connections, and the parameters of the sigmoid function. Determining the subset of C1 subunits to connect to an S2 unit is an NP-complete problem, and we chose the heuristic based, forward selection algorithm, greedy search to find a solution (Russell and Norvig 2003). Although we could have applied other methods for solving NP-complete problems, we chose greedy search for its simplicity and for its efficacy in this problem domain. Figure 2 shows an overview schematic of the forward selection fitting procedure. The search was initialized by evaluating all possible combinations of two subunits taken from the 3×3 C1 grid size. At each step within the search we determined the parameters for each C1 subunit combination using gradient descent in parameter space, which included the C1 weights and the sigmoid parameters, to minimize the mean squared error between the experimentally measured V4 response and the C2 unit's response (note that under a probabilistic interpretation, minimizing the mean squared error implies a Gaussian noise distribution around the measured responses). Within each iteration step of the greedy search, the combination of n C1 units producing lowest mean squared error between the experimental V4 measurements, and the model responses was selected as the winner. In the next iteration step the algorithm searched over every possible combination of $n + 1$ C1 units to find a better fit (the winning configuration from the previous iteration plus an additional C1 unit not previously selected).

Depending on the aspect of the model we wished to analyze, we determined the number of C1 subunits by one of two methods. The first method was used to find a single model for each V4 neuron (as in Figs. 3 and 4) and used cross-validation to mitigate overfitting. In this method, the number of subunits was set to the minimum number of units, between 2 and 25, that minimized the average testing error over a sixfold cross-validation set to within 1% of the absolute minimum. An n-fold cross-validation divides the dataset into n equal-sized randomly selected subsets, trains the model on $n - 1$ of the subsets, and predicts the response on the remaining subset. This is repeated n times, each time predicting a different subset (see supplemental materials figure S1[1] for an example of the training and testing errors as a function of the number of subunits for each fold). Subsequently, the best fitting C2 unit with this number of subunits was found over the entire dataset. For each C2 unit, we limited the maximum number of subunits to 25.

[1] The online version of this article contains supplemental data.

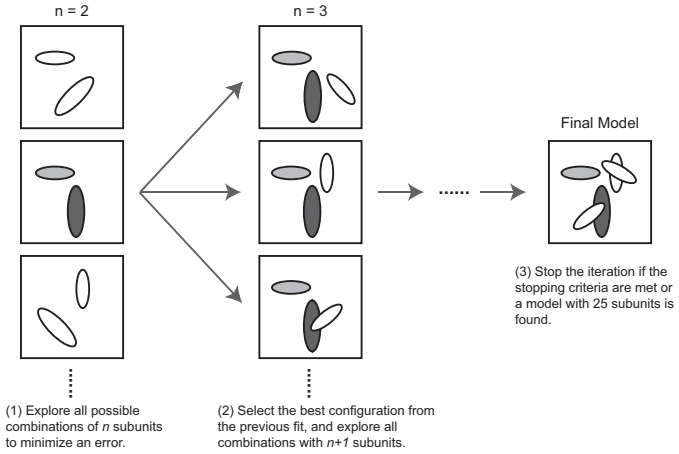

FIG. 2. Schematic of the model fitting procedure. The response of each V4 neuron was fit with a model C2 unit by determining the parameters between the C1 layer and the S2 layer (see box in Fig. 1, parameter fitting). Because the response of C2 units is highly nonlinear and an analytic solution is intractable, we used numerical methods to find solutions. For each model fit, we determined the set of C1 subunits connected to the S2 units, the weights of the connections, and the parameters of the sigmoid function. To determine the subset of C1 subunits, we used a forward selection algorithm, greedy search, to find a solution. The search was initialized (*left*) by selecting 2 C1 subunits to include in the selectivity function. For each selection of 2 subunits, the parameters of the selectivity function were adjusted using gradient descent in parameter space to minimize the mean squared error between the V4 neuron's measured response and the model C2 unit's response. The C1 subunit configuration that achieved the lowest mean squared error was then used for the next iteration of the greedy search. The search then continued (*middle*) by adding an additional C1 subunit to the best configuration found in the previous iteration. The search was stopped (*right*) to produce a final model. One of 2 stopping criteria was chosen based on the desired analysis of the final model. To find a single model for each V4 neuron, the search was halted once the average testing error over a sixfold cross-validation set reached within 1% of the absolute minimum. To test the model's ability to generalize to stimuli outside the training set, the search was stopped once the mean squared error on the training set decreased by $<1\%$ or once 25 C1 subunits were found. See *Fitting model parameters* for further discussion.

To test the model's ability to generalize to stimuli outside the training set, we used a second method for determining the number of C1 subunits found in the fitting procedure. We split the stimulus set into randomly selected training and testing sets containing 305 and 61 stimulus-response pairs, respectively. The number of C1 subunits was determined on the training set by adding subunits in the greedy search until the error between the C2 unit's response and the V4 neuron's response decreased by $<1\%$ or once 25 C1 subunits were found. We then simulated the resulting C2 unit's response on the test set, measuring the model's ability to generalize to stimuli outside the training set (as in Fig. 5). This method is often referred to as validation.

In summary, the model of a V4 cell is derived from a parameter space consisting of 119 free parameters (synaptic weights for 116 C1 units and 3 sigmoid parameters). However, for fitting the responses of each V4 neuron over 366 stimuli, a small subset of these parameters is selected based on cross-validation criteria. Over the population of V4 neurons examined, the median number of parameters chosen was 13, with a minimum of 5 ($2 + 3$) and a maximum of 28 ($25 + 3$). Notice that even with a large number of possible parameters our model is still highly constrained by its structure.

RESULTS

Selectivity for boundary conformation

C2 units in the model can reproduce the selectivity of V4 neuronal responses. Model neurons reproduce the variety of

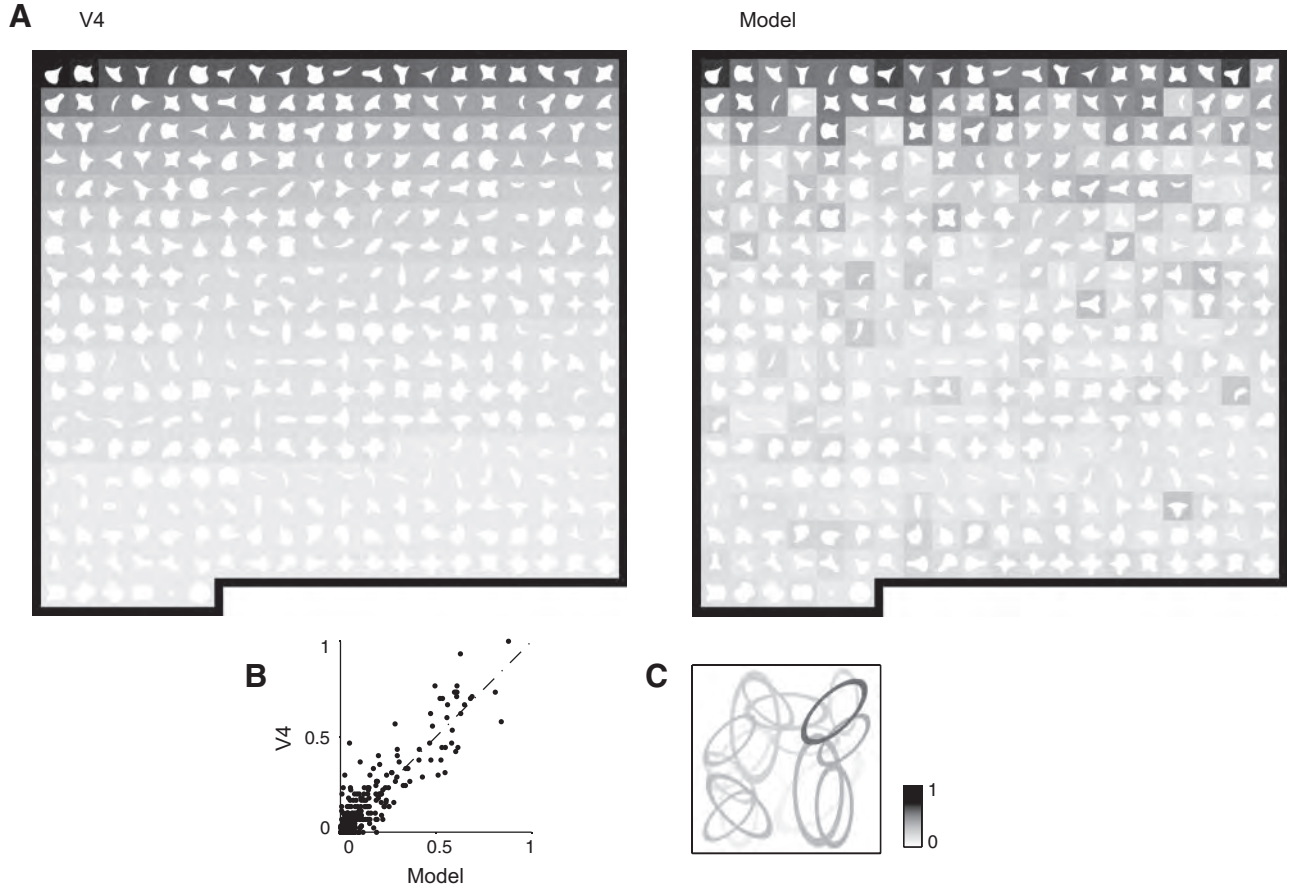

FIG. 3. Comparison of model responses to a V4 neuron tuned to convex curvature. *A, left*: the selectivity of a V4 neuron over 366 boundary conformation stimuli is shown in order of decreasing response strength. The magnitude of the response is indicated by the gray scale (high response is darker). From the inspection of the response profile, it is apparent that this neuron is selective for a high convexity, or a sharp angle protruding out, on the upper right side of a stimulus. This is the same neuron that appears in Fig. 5 of Pasupathy and Connor (2001). *Right*: response of the C2 unit, modeling this V4 neuron's response, shown in the same stimulus order. A similar selectivity profile is observed. *B*: response of the V4 neuron is plotted against the model C2 unit's response for each stimulus. The goodness of fit, measured by the correlation coefficient, is 0.91 between this neuron and the model over the 366 boundary conformation stimuli. *C*: configuration of C1 subunits, projecting to S2 model units, is shown schematically. The configuration and weights of C1 afferents determine the selectivity of the S2 units and the resulting C2 unit. The locations and orientations of the C1 subunits are indicated by ellipses, and the strength of the synaptic weight is indicated by gray scale. This particular C2 unit is composed of S2 units each of which combines 18 C1 subunits with 1 strong afferent pointing diagonally outward in the upper right corner of the receptive field. This configuration is typical of C2 units that produce tuning to sharp curvature projections within the stimulus space.

selectivity described previously in V4 (Pasupathy and Connor 2001), including selectivity to angular position and the curvature of boundary fragments. Figure 3 compares the responses of an example V4 neuron to the corresponding C2 unit. This V4 neuron is selective for sharp convex boundary fragments positioned near the upper right corner of a stimulus, as shown in the response-magnitude ranked illustration of the stimuli in Fig. 3*A*. The modeled responses correspond closely to the physiological responses (coefficient of correlation $r = 0.91$, explained variance $r^2 = 83\%$; note that fitting V4 neural selectivity with a C2 unit is a more difficult problem than fitting selectivity at the S2 level because the invariance operation, or pooling, of the C2 unit may cause interference between the selectivities of translated S2 units). This type of selectivity is achieved by a S2 configuration with 18 C1 subunits, shown schematically in Fig. 3*C*, which form a nonlinear template for the critical boundary fragments. The configuration of the C1 subunits offers a straightforward explanation for the observed selectivity. The C2 unit has a C1 subunit at 45° with a high weight, oriented along the radial direction (also at 45°) with respect to the center of the receptive field. This subunit configuration results in selectivity for sharp projections at 45° within the stimulus set and is described by the boundary conformation model as tuning for high curvature at 45° relative to the object center (see *Comparison with the curvature and angular position tuning model* for an analysis of the correspondence between C1 configurations and curvature tuning).

C2 units can also reproduce selectivity for concave boundary fragments. Responses of the second example neuron, Fig. 4, exhibit selectivity for concave curvatures in the lower part of a stimulus. Again, there is a strong correspondence between the modeled and measured responses ($r = 0.91$, explained variance = 83%). In this example, selectivity was achieved by a S2 configuration with 23 oriented subunits, shown schematically in Fig. 4*C*. Note that there are several separated subunits with strong synaptic weights in the lower portion of the receptive field at −45, 0, and 45° orientations; these correspond to boundary fragments found in many of the preferred stimuli. In general, the geometric configuration of oriented subunits in the

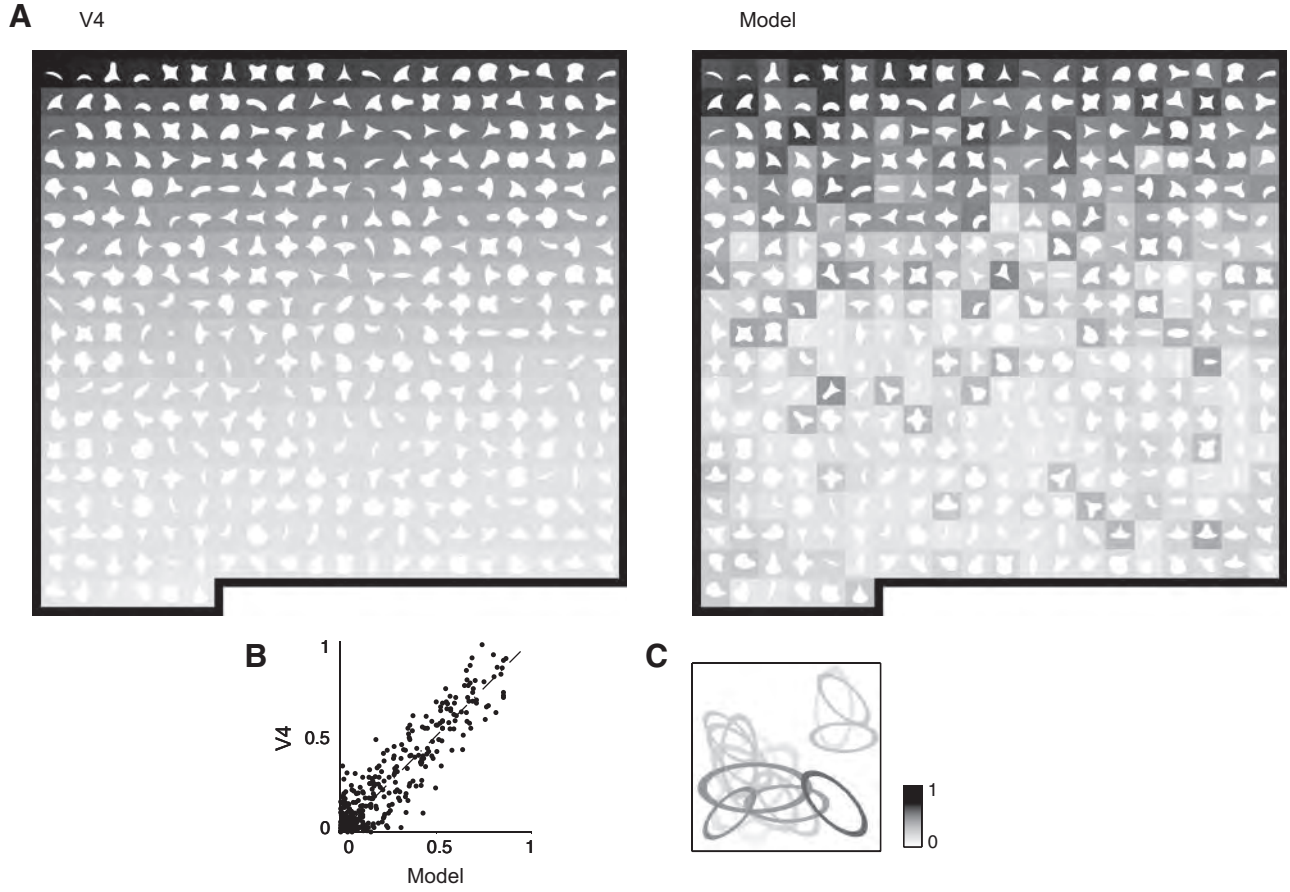

FIG. 4. Comparison of model responses to a V4 neuron tuned to concave curvature. The selectivity of another example neuron in the same format as Fig. 3 is shown. *A*: this V4 neuron shows selectivity to boundary conformations with slightly concave curvature, or an obtuse angle, in the lower portion of the receptive field. *B*: model C2 unit closely matches the V4 neuron's response ($r = 0.91$). *C*: S2 configuration of the model is quite complex with 21 afferent C1 subunits. The group of dominant subunits, oriented at 45, 0, and −45° in the lower portion of the S2 receptive field, has a strong influence on the observed selectivity.

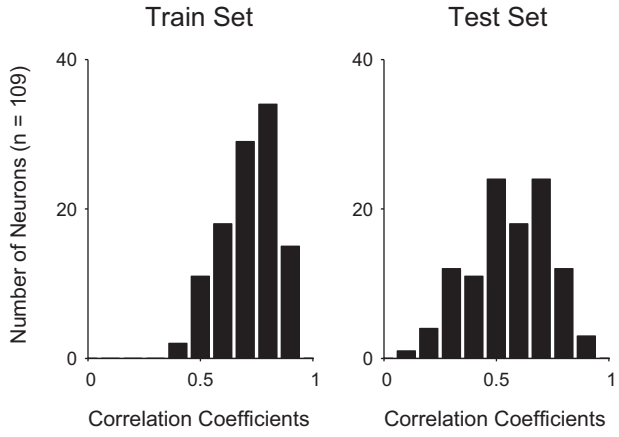

FIG. 5. Generalization of the model to stimuli outside of the training set. The model is able to predict the response of V4 neurons to boundary conformation stimuli not included in the training set. Using a sixfold cross-validation methodology across the population of V4 neurons, a model C2 unit was determined for each V4 neuron using a training set, and the resulting model was used to predict the V4 neuron's response to the testing set. A histogram of correlation coefficients on the training (*left*) and testing (*right*) sets are shown. Over the population, the median correlation coefficients were 0.72 on the training set and 0.57 on the testing set. These numbers are based on the averages over the sixfold cross-validation.

model closely resembles the shape of a critical region in the stimuli that elicit high responses.

Testing population selectivity for boundary conformation

Model C2 units can successfully fit the V4 population selectivity data and can generalize to V4 responses outside the training set. For each V4 neuron, we divided the main stimulus set randomly into two nonoverlapping groups (a training and a testing set) in a standard cross-validation procedure (see METHODS). Figure 5 shows correlation coefficient histograms for training and testing over the population of V4 neurons. The median correlation coefficient between the neural data and the C2 unit responses was 0.72 (explained variance = 52%) on the training set, and 0.57 (explained variance = 32%) on the test set over sixfold cross-validation splits of the dataset. However, because the stimulus set is inevitably correlated, the test set correlation coefficients are inflated. The full distributions of the model parameters can be found in supplemental figure S2.

Much of the variance in V4 neuron responses may be unexplainable due to noise or uncontrolled factors. Pasupathy and Connor (2001) estimated the noise variance by calculating the average expected squared differences across stimulus presentations. The estimated noise variance averaged 41.6% of the total variance. Using this estimate, on the training set the model

accounted for 89% of the explainable variance ($r = 0.94$) and on the testing set the model accounted for 56% of the explainable variance ($r = 0.75$). Therefore a large part of the explainable variance is described by the model. This result indicates that the model can generalize within the boundary conformation stimulus set.

Invariance to translation

The model not only matches V4 selectivity but also reproduces V4 translation invariance. Responses of V4 neurons are invariant to translation (i.e., their selectivity is preserved over a local translation range) as reported in many studies (Desimone and Schein 1987; Gallant et al. 1996; Pasupathy and Connor 1999, 2001). The population of C2 units used to fit the population of V4 neurons reproduced the selectivity of those V4 neurons, while still maintaining invariance to translation. Selectivity and invariance are two competing requirements and the model C2 units satisfy both requirements. The results in Fig. 6 show that the built-in invariance mechanism (at the level of C2) operates as expected, reproducing the observed translation invariance in the experimental data on the boundary conformation stimuli. Figure 6*A* shows the invariance properties of the C2 unit from Fig. 3. Eight stimuli, which span the response range, are sampled across a 5 × 5 grid of positions with intervals equal to half the classical receptive field radius. Not only does the stimulus that produces a high response at the center of the receptive field produce high responses over a range of translation, but more importantly, the selectivity is preserved over translation (i.e., the ranking of the eight stimuli is preserved over translation within a given range). Figure 6, *B* and *C,* shows that the observed translation invariance of V4 neurons is captured by the population of C2 units. Because the C2 units are selective for complex, nonlinear conjunctions of oriented features and the invariance operation is based on pooling from a discrete number of afferents, the translated stimuli sometimes result in changes of selectivity. A few C2 units in Fig. 6*B* show that translated nonoptimal stimuli can produce greater responses; but on average, as shown in Fig. 6*C*, optimal stimuli within a range of translation produce stronger responses.

Responses to bar and grating stimuli

The model is capable of reproducing the responses of individual V4 neurons to stimuli not determined by boundary conformation, such as bars and gratings. The population of C2 units produces responses that are consistent with the general findings that populations of V4 neurons show a wide range of orientation selectivity and bandwidths, individual V4 neurons exhibit multiple peaks in their orientation tuning curves, and V4 neurons show a strong preference for polar and hyperbolic gratings over Cartesian gratings.

To compute the orientation bandwidth of each C2 unit, the orientation selectivity of each model unit was measured using bar stimuli at various orientations (10° steps), widths (5, 10, 20, 30, and 50% of the receptive field size), and locations within the receptive field. The orientation bandwidth of each model C2 unit, the full width at half-maximum response, with linear interpolation as in Fig. 6*A* of Desimone and Schein (1987), was taken for the bar that produced the highest re-

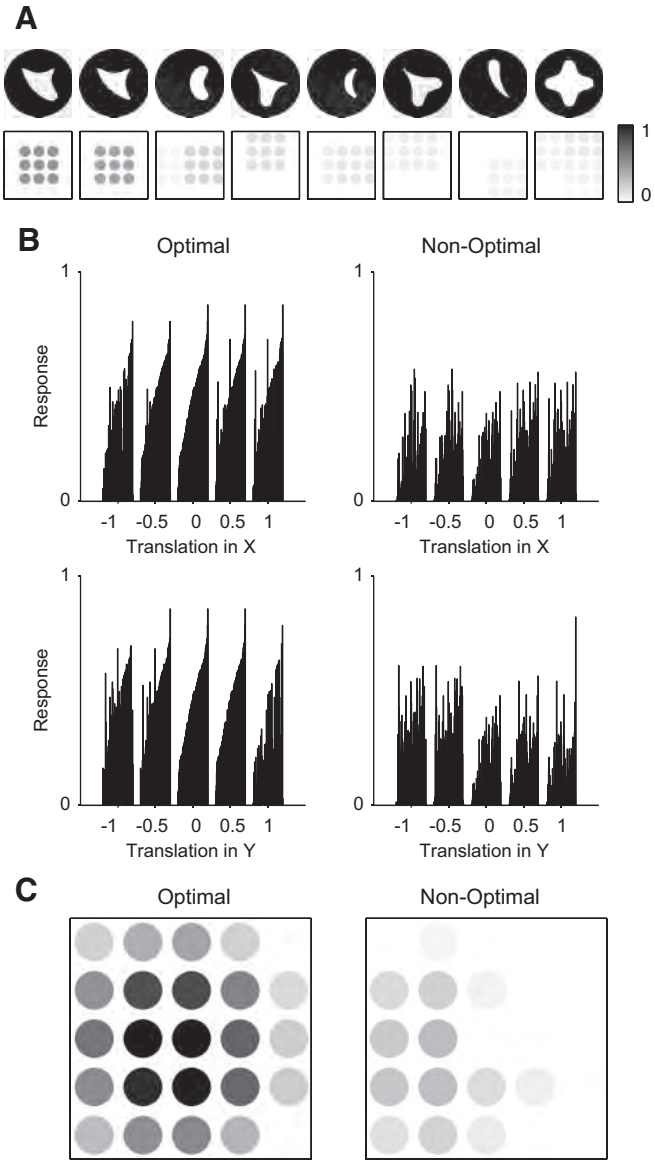

FIG. 6. Translation invariance of model neurons. C2 units are invariant to translations, comparable to the invariance observed in V4. *A, top*: 8 different stimuli, which elicit a wide range of responses from the C2 unit of Fig. 3, are shown; *bottom*: corresponding responses of this C2 unit are shown. Each stimulus was presented at 25 positions on a 5 × 5 grid (separated by half an S2 receptive field radius) centered on the C2 receptive field, following Fig. 6*A* of Pasupathy and Connor (2001). The center position in the 5 × 5 grid corresponds to the default presentation condition. Note that selectivity is preserved over a range of translation. *B*: population of 109 C2 units also shows invariance to translation and preserved selectivity over translation. The responses of the 109 C2 units to the stimulus that produced the highest response (optimal) and the stimulus that produced the lowest response (nonoptimal) from the same set of 8 stimuli from *A* are displayed. Each thin bar corresponds to the response of a C2 unit, averaged along the orthogonal directions within the 5 × 5 grid, and the bars are sorted according to the responses at the default presentation condition. The *x* axes are in units of S2 receptive field size. *C*: this figure shows the average, normalized responses to the stimuli that produced the highest (optimal) and lowest (nonoptimal) responses out of the 8 shown in the *top row* of *A*, across 109 C2 units for each stimulus position. The selectivity is generally preserved over translation. The responses to the optimal and nonoptimal stimuli at the central position for each C2 unit are normalized to be 1 and 0, respectively, so that each unit makes equal contribution to this plot.

sponse across location and orientation. The multimodal nature of the orientation tuning curves was assessed using a bimodal tuning index, Eq. 9 in David et al. (2006). To find the bimodal tuning index, we first found the two largest peaks and the two smallest troughs in the orientation tuning curve. The index is computed by taking the ratio of the difference between the smaller peak and larger trough, to the difference between the larger peak and smaller trough. Orientation tuning curves with only one peak have an index value of 0 and orientation tuning curves with tuning peaks and troughs of equal size will have a bimodal tuning index of 1.

Figure 7*A* provides a summary plot of orientation bandwidths measured for 97 model C2 units (of 109 C2 units, 97 had a response to a bar stimulus that was ≥10% of the maximum response to the contour stimulus set). The distribution of orientation bandwidths covers a wide range that is comparable to the physiologically measured range from Desimone and Schein (1987) and David et al. (2006). The median orientation bandwidth for the C2 population was 51.7°, whereas the median found in Desimone and Schein (1987) and David et al. (2006) was around 74°. The larger median orientation bandwidth in the physiological measurements is a product of the large portion of V4 cells found to be nonorientation selective in the physiological population [32.5% of cells in Desimone and Schein (1987) had orientation bandwidths >90°] and the small portion of C2 units found with a similar lack of orientation selectivity (~8% of C2 fits had orientation bandwidths >90°). When only considering V4 cells with orientation bandwidths <90°, Desimone and Schein (1987) found that the median orientation bandwidth was 52°, similar to the median of 51.7° over C2 fits. This discrepancy between the two populations may be due to a selection bias in the recordings of Pasupathy and Connor (2001), who selected cells based on their tuning to complex shapes. For this reason, cells with a lack of selectivity, those with orientation bandwidths >90°, may not have been included in their recordings.

Individual orientation tuning curves also indicated that many model C2 units were selective for multiple orientations. Such multi-modal orientation tuning, as opposed to the unimodal tuning in V1, is one of the characteristics of V4 neurons, and it arises naturally in our model because each model unit is composed of several differently oriented subunits. Although a number of model units have more than two peaks in their tuning curves, we computed bimodal tuning indices, which characterize the deviation from the unimodal tuning behavior (David et al. 2006). Figure 7*B* presents a summary plot of bimodal tuning index for the 97 model C2 units that were responsive to the bar stimuli. The overall range of the bimodal index distribution in Fig. 7*B* is comparable with Fig. 5*D* in David et al. (2006) and has a similar profile: a peak near zero with a dramatic falloff as the bimodal index increases. The median bimodal index over the population of C2 units was 0.12 and over the V4 population measured in David et al. (2006) the median bimodal index was 0.09.

To test individual C2 units to grating stimuli, we used the same 109 model C2 units fit to the V4 population and presented three types of gratings: 30 Cartesian, 40 polar, and 20 hyperbolic gratings each at four different phases to reproduce the stimulus set used in Gallant et al. (1996). The boundary conformation stimuli produced an average response of 0.22 from 109 C2 units, whereas the polar and hyperbolic grating

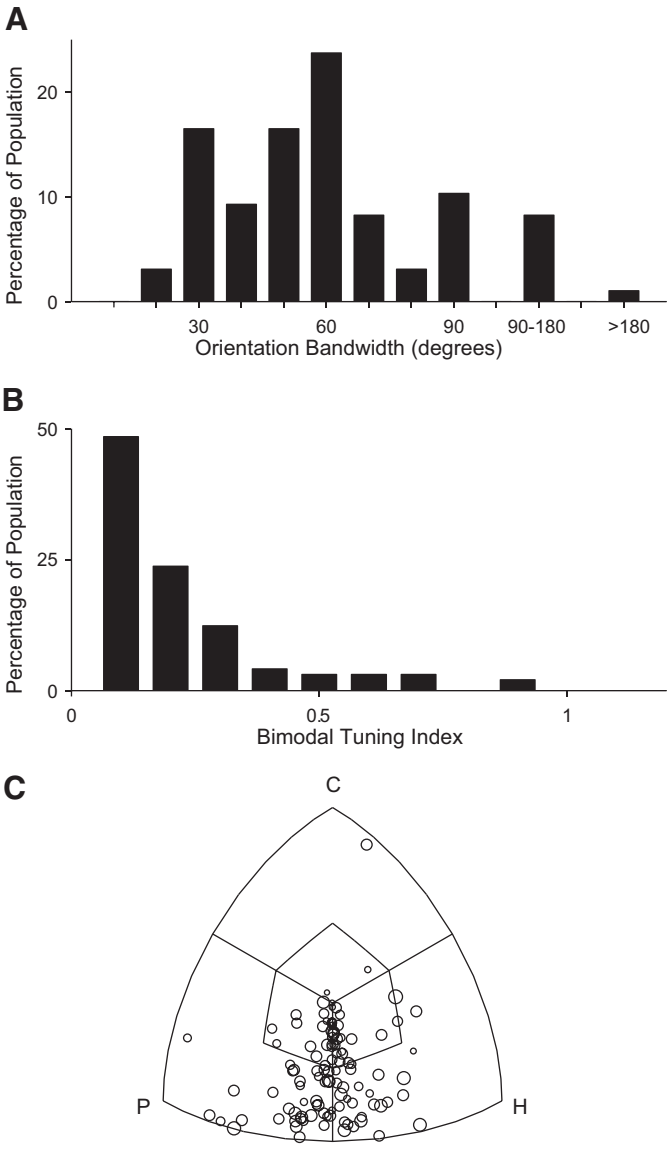

FIG. 7. Testing selectivity of C2 units to bars and gratings. We measured the responses from a population of C2 units fit to the population of V4 neurons from Pasupathy and Connor (2001), using bars and gratings. Three summary plots are presented: *A*: orientation bandwidth; *B*: bimodal tuning index; and *C*: tuning to Cartesian, polar, and hyperbolic gratings. *A* shows a histogram of orientation bandwidths measured for 97 of the C2 units that showed significant response to bar stimuli. The median orientation bandwidth for the C2 population was 51.7°. *B* shows a histogram over bimodal tuning index for the same 97 model C2 units. The median bimodal indexes over the population of C2 units is 0.12. *C* shows a summary plot of all 109 model C2 units to Cartesian, polar, and hyperbolic gratings. The grating stimuli, analysis procedure, and plotting convention used in Gallant et al. (1996) are reproduced to asses the selectivity to complex gratings. For each C2 unit, the maximum responses to each grating class (Cartesian, polar, and hyperbolic) form a 3-dimensional vector, normalized to unit length, and plotted in a 3-dimensional space with each axis representing the response to a grating class (the viewpoint is oriented so that the origin of the coordinate system is at the center, and the vector whose responses are equal is pointing directly out of the page). The vector for each model C2 unit is plotted as a circle with the size of the circle indicating the magnitude of the highest response over all of the grating stimuli. The bias toward polar and hyperbolic gratings, which is a characteristic previously described in V4 (Gallant et al. 1996), indicates that for most C2 units, the optimal stimulus was non-Cartesian. Our results show a stronger bias than reported in (Gallant et al. 1996).

stimuli produced an average response of 0.14 (1.0 is the maximum measured response over the main boundary conformation stimulus set). However, for 39% of the C2 units, the most preferred stimulus was one of the grating stimuli and not one of the boundary conformation stimuli. This result suggests that some V4 neurons selective for curved object boundary fragments might also show significantly higher responses to grating stimuli and other complex patterns.

In correspondence with the report of a distinct group of V4 neurons that are highly selective for hyperbolic gratings (Gallant et al. 1996), we also found individual C2 units within our population highly selective for hyperbolic gratings. For example the C2 unit used to model the V4 neuron in Fig. 3 showed a strong preference for hyperbolic gratings, as its maximum response over hyperbolic gratings, 0.90, was much greater than the maximum responses over both polar gratings, 0.39, and Cartesian gratings, 0.04.

The population of C2 units also reproduces previously measured V4 population response characteristics to gratings. The distribution of grating class selectivity is shown in Fig. 7*C*. Quantitatively, mean responses to the preferred stimulus within each grating class were 0.004 for Cartesian, 0.160 for polar, and 0.196 for hyperbolic, qualitatively matching the finding in Gallant et al. (1996) that the population of V4 neurons they measured is strongly biased toward non-Cartesian gratings. Many of the C2 units produced a maximal response to one grating class at least twice that of the other two classes: 1% for Cartesian, 35% for polar, and 26% for hyperbolic gratings. The reported experimental findings were 2, 11, and 10%, respectively.

The C2 population tends to be more strongly responsive to the non-Cartesian gratings than reported in Gallant et al. (1996). This discrepancy may be due to different screening processes used in the two experiments [V4 neurons in Pasupathy and Connor (2001) were recorded only if they responded to complex stimuli, and were skipped if they appeared responsive only to bar orientation]. The C2 population also tends to show less-selective responses between the polar and hyperbolic gratings than the neural data as indicated by the concentrated points near the polar-hyperbolic grating boundary in Fig. 7*C*. An earlier modeling study (Kouh and Riesenhuber 2003) suggests that a larger distance between the orientation-selective subunits can increase the variance of responses to these non-Cartesian grating classes, but this parameter was fixed in all of our simulations.

Model architecture, complexity and limitations

TWO-LAYER MODEL ARCHITECTURE. Our model of V4, as shown in Fig. 1, uses a C2 layer to explicitly implement translation invariance and localized S2 units to achieve selectivity. Such a construct is a consistent part of a canonical architectural principle of the full model of the ventral pathway (Fig. 1), aimed at gradually building up selectivity and invariance for robust object recognition. Could S2 units, receiving input directly from complex V1-like neurons, reproduce both selectivity and invariance exhibited by V4 neurons? To test this hypothesis, responses to a stimulus set derived from the measurements of a V4 neuron that tested both the neuron's selectivity and invariance were fit with four different models: *C2 unit*, the full C2 unit implementation that pools locally selective S2 units; *S2 unit*, a single S2 unit identical to those used in the full C2 model; *Control 1*, a single S2 unit modified to receive inputs from spatially localized C1 units collectively spanning the receptive field of the V4 neuron; and *Control 2*, a single S2 unit modified to receive input from nonspatially localized C1 units that achieved translation invariance over the entire V4 receptive field. For control 1, the population of C1 units included the entire population of C1 units used in the full C2 model. For control 2, the population of C1 units was created by performing the invariance operation (maximum operation) over C1 units spanning the entire receptive field with identical orientation and bandwidth.

Each model was evaluated on a stimulus set that tested both selectivity and invariance. A 5×5 translation grid, with each stimulus translated by 50% of the classical receptive field radius and identical to that used in Pasupathy and Connor (2001), was used to create a total of 9,150 stimuli (main stimulus set $\times$ 25 translated positions). The corresponding V4 response to all these stimuli was derived from the selectivity response of a single V4 neuron by replicating the response to the centered stimulus over a grid matching translation invariance range typical of the population of V4 neurons (in this case the central 3×3 grid). Note that this represents an idealized response set and actual V4 responses are slightly more varied over translation, see Fig. 6*A* from Pasupathy and Connor (2001). Each model was fit to this stimulus set using the same cross-validation fitting procedure described in *Fitting model parameters* within METHODS. This allowed us to quantitatively measure the selectivity and invariance of each model using a correlation coefficient on the testing set. We also qualitatively assessed the degree of translation invariance for each model.

The C2 unit was the model that best matched V4 selectivity and invariance. For each cross-validation fold, we computed the correlation coefficient on the testing set for each model as a function of the number of subunits, shown in Fig. 8*A*. Clearly, the C2 unit reaches a higher correlation coefficient than the other models and produces better fits over the range of subunits tested. The test set correlation coefficient averaged over the cross-validation folds (using the subunit stopping criteria of training error decreasing by $<1\%$) for the C2 unit was 0.79 ± 0.014 (mean $\pm$ SD; explainable variance = 62%), whereas the correlation coefficients for the S2 unit, control 1, and control 2 were 0.61 ± 0.022 (37%), 0.65 ± 0.015 (42%), and 0.35 ± 0.013 (12%), respectively. For this stopping criterion, the average number of subunits for each model was 16.0, 7.2, 10.3, and 2.7, for the C2 unit, the S2 unit, control 1, and control 2, respectively.

A qualitative demonstration of translation invariance also shows that only the C2 unit maintains selectivity over translation. Figure 8*B* shows four stimuli that span the range of the V4 neuron's response (*1st column*), the derived V4 response profile for each stimulus over the 5×5 translation grid (*2nd column*), and the response of each model to the same stimuli (*remaining columns*). The derived V4 response shows translation invariance over a limited range (the central 3×3 portion of the translation grid) and maintains the selectivity profile over translation (i.e., for each translation the stimulus ranking is preserved). Only the C2 unit shows both the required range of translation invariance and maintains the stimulus ranking. The S2 unit produces a high degree of variation across translation and fails to maintain the stimulus ranking. Control 1

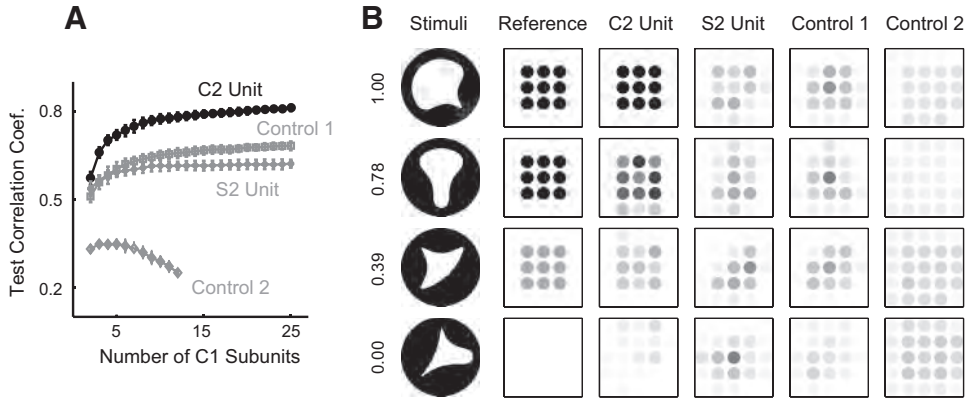

FIG. 8. Comparison of model architectures for selectivity and invariance. Our V4 model architecture, which consists of a C2 layer to explicitly implement translation invariance and a layer of localized S2 units to achieve selectivity, is necessary to reproduce both the selectivity and invariance characteristics of V4. We compared the selectivity and invariance properties of 4 model architectures: C2 unit, the full C2 unit implementation; S2 unit, a single S2 unit; control 1, a single S2 unit modified to receive inputs from spatially localized C1 units; and control 2, a single S2 unit modified to receive inputs from nonspatially localized C1 units. *A*: plots the correlation coefficients between the cross-validation testing set and each model's response as a function of the number of C1 subunits (see TWO-LAYER MODEL ARCHITECTURE for simulation details). *B,* first column: 4 stimuli that span the range of the V4 neuron's response (normalized numerical values are indicated by the numbers to the left of each stimulus). *Second column*: simulated response profile to each stimulus over the 5×5 translation grid. *Remaining columns*: fit response of each model to the same stimuli. Note that the derived V4 response shows translation invariance over a limited range and maintains the selectivity profile over translation (for each translation the stimulus ranking is preserved). Only the C2 unit shows both the required range of translation selectivity and maintains the stimulus ranking.

does reproduce the stimulus ranking over the central translation position, but fails to achieve translation invariance. Control 2's response maintains a high degree of translation invariance (the underlying C1 population response is invariant to translation), but it does not reproduce the stimulus ranking for any of the translated positions.

These controls provide justification for our model architecture of a two layer S2–C2 hierarchy to produce both selectivity and invariance that matches the observed responses in V4. Selectivity and invariance are in general competing requirements that are difficult to satisfy at the same time (Mel and Fiser 2000). Therefore in our model, they are gradually built up in alternating layers with separate operations for selectivity and invariance. For V4, spatially localized selectivity units (S2 units) are pooled over position by C2 units to achieve selectivity and invariance. This is one of the main computational principles of our model of the ventral pathway (Fig. 1) (see Riesenhuber and Poggio 1999; Serre et al. 2005). These control experiments suggest that this mechanism may play a central role in the computations performed by V4 neurons.

COMPLEXITY OF V4 NEURONS. Based on our model, we sought to estimate the complexity of the V4 neuron population. Figure 9*A* shows a distribution of the number of C1 afferent units found by the cross-validation analysis (see METHODS). The results for predicting stimuli outside the training set, Fig. 5, are based on this distribution of C1 subunits. The median number of C1 afferent units found for the distribution was 13. In other words, a median of 16 parameters (13 plus 3 parameters in the sigmoid function, *Eq. 2*) were required to explain the measured V4 responses to the boundary conformation stimulus set. Figure 9*B* shows the evolution of the correlation coefficients of the predicted responses for each V4 neuron and their mean over the neurons. The mean correlation coefficient for a given number of C1 afferents continues to improve all the way up to 25 C1 afferents. There was a significant correlation of 0.47 ($P < 0.001$) between the mean correlation coefficient and the number of C1 afferents (see supplemental figure S3). This indicates that adding additional C1 afferents according to our methodology does not result in overfitting of the neural responses and, within the framework of our model, that these additional C1 afferents are necessary for estimating the complexity of V4 neurons. Our model predicts that V4 neurons are not homogeneous in their complexity, but span a continuum in their selectivity to complex stimuli. This continuum is illustrated by the S2 configuration diagrams of all 109 neurons in Fig. 9*C*.

INCREASING THE PARAMETER SPACE. We tried to determine if the overall conclusions drawn from our model were highly dependent on the underlying S1–C1 hierarchy. While we did find that it was possible to produce model C2 units with fewer C1 afferents if the parameter space was more densely sampled, we found that the conclusions of our analysis remained unchanged. To demonstrate these points, we increased the number of S1 orientations from four to eight and placed a stricter limit on the number of subunits (at most 10 C1 afferents per S2 unit). The resulting model achieves a similar level of fit to the boundary conformation stimulus set with correlation coefficients of 0.71 (explainable variance = 50%) on the training set and 0.56 (31%) on test set (see supplemental material Fig. S4 and compare with Fig. 5). The median number of C1 units used for the fitting was eight (cf. 13 in Fig. 9*A*). The results were also similar for translation invariance, responses to the grating stimuli, and comparisons with the curvature model (qualitatively identical to Figs. 6, 7, and 10). The geometric configurations with the extended C1 subunit types also show close resemblance to the previous results. For instance, the new configuration for the C2 unit in Fig. 3 is still composed of a radially oriented subunit in the upper right corner, and the C2 unit in Fig. 4 is still composed of two widely separated C1 subunits in the lower part of the receptive field as shown in supplemental material Fig. S4*A*. Because the number and arrangement of subunits were largely unaffected, we concluded that altering the complexity of the C1 layer would not affect the descriptive power or conclusions of our model.

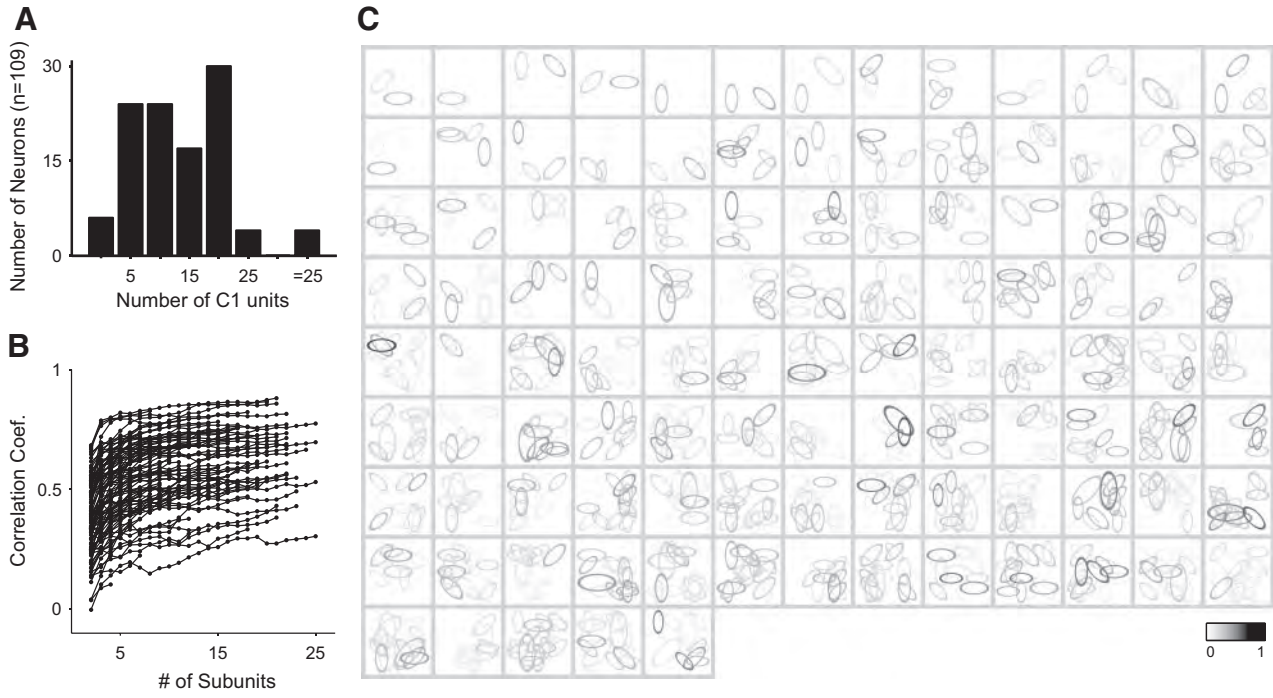

FIG. 9. Complexity distribution of C2 units. *A*: varying number of afferent subunits is required to fit the 109 V4 neurons. Some C2 units had only a few subunits, while others required >20. The median was 13. The number of subunits was chosen to achieve the minimum average test error over sixfold cross-validation (see METHODS). A maximum of 25 subunits was used in the fitting. *B*: evolution of the correlation coefficients (on the test set of the cross-validation) is shown along with the corresponding number of C1 afferent units. Individual C2 units are represented by each line, and based on the cross-validation criteria, fewer than 25 subunits are used for the final fit. Models with 10 subunits produced an average correlation coefficient of 0.5 between the model and the neural responses. *C*: S2 configurations for all 109 C2 units are shown in order of increasing number of afferents, illustrating that a population of V4 neurons is likely to come in widely different configurations and combinations of afferent inputs.

LIMITATIONS OF THE MODELING FRAMEWORK. The specific C2 model units described here fail to capture some known aspects of ventral stream visual responses. In general, the C2 model units cannot easily be used to describe invariances or selectivities that are much more complex than those seen in V4. For example, our present model C2 units could not capture the degree of selectivity and invariance found in the responses of neurons in inferotemporal cortex. Scale invariance is another characteristic of visual processing that is not easily captured by the currently formulated C2 units. However, it should be possible to build a similar pooling mechanism over S2 units of different scales to achieve scale invariance at the C2 level (Riesenhuber and Poggio 1999). A detailed study of scale invariance within V4 would provide additional constraints on subsequent models.

In addition the model captures only a fraction of the response variance in a portion of the V4 population we have analyzed. We could not determine any clear pattern among the responses of neurons that were fit poorly by the model. Whereas these poor fits may be due to noise variance or distinct functional populations of neurons within V4, they may also represent a fundamental limitation of our model. Given the current V4 data, it is unclear if nonlinear feedforward models of this type will fundamentally fail at explaining initial V4 responses (without attentional modulation). To achieve a more detailed understanding of V4, it will be necessary to use stimuli that push the limits of known models. Taken together, these limitations indicate that the current data on V4 do not provide a clear distinction between the functional operation of V4 and the model of visual processing we have described.

LIMITATIONS OF THE CURRENT FITTING FRAMEWORK. One of the main limitations of the current fitting framework is the stability of the solution. In other words, for a given response profile of a V4 neuron, the geometric configuration of the C1 subunits, obtained by the fitting procedure, is underconstrained and not guaranteed to be unique because there exist other configurations that would yield a similar level of fit with the neural response. However, most fitting results converged onto similar geometric configurations (compare the configurations in Figs. 3*C* and 4C, with supplemental Fig. S4*A*). Regardless of the exact solutions, our modeling approach provides an existence proof that a model based on combining spatially localized selectivity units can account for V4 tuning data. Our approach does not require uniqueness, as finding several afferent combinations that all can account for the experimentally observed tuning and invariance data lead to this same conclusion.

COMPARISON WITH THE CURVATURE AND ANGULAR POSITION TUNING MODEL One goal of our model is to understand how curvature and angular position tuning could be achieved from the known representations in lower visual areas. C2 units provide a mechanistic explanation of V4 selectivity, whereas in Pasupathy and Connor (2001), tuning functions on curvature and angular position of the boundary fragments provide another description of the response profiles of the recorded V4 neurons. Therefore we examined the correspondence between the configurations of S2 afferents with the tuning functions for curvature and angular position derived in Pasupathy and Connor (2001). We compared C2 model fits with three aspects of the 4D curvature and angular position tuning functions de-

scribed in (Pasupathy and Connor 2001): the goodness of fit (correlation coefficient), the peak locations of angular position, and the degree of curvature.

Both C2 units and 4D curvature-angular position tuning functions capture much of the response variance of V4 neurons. The median training set correlation coefficients of the 2D and 4D curvature-angular position tuning models were 0.46 and 0.57, respectively (see Pasupathy and Connor 2001 for a description of these models). There is a high correspondence between the correlation coefficients found for C2 units and the curvature-angular position tuning fits (shown in Fig. 10*A*). This may not be surprising, as both models produce tuning functions in the space of contour segments that make up these stimuli.

We investigated the correspondence between the curvature-angular position tuning and the parameters of model C2 units. In many cases, there is an intuitive relationship between the geometric configuration of a C2 unit's oriented C1 afferents and the tuning parameters in curvature and angular position space (i.e., Fig. 3*C*, concave curvature tuning, and Fig. 4*C*, convex curvature tuning, show such correspondence at specific angular positions). To quantitatively examine this relationship, we examined the parameters at the S2 level and compared them to the peak locations of angular position and the degree of curvature found with the parameterized tuning functions. We found that angular position tuning is closely related to the weighted average of subunit locations, illustrated in Fig. 10*B*. Because the receptive fields of S2 units are large in comparison to C2 units (S2 RF radius = 0.75 × C2 RF radius), any spatial bias in the C1 inputs to S2 units will create a spatial bias at the C2 level. If this spatial bias is concentrated, the C2 unit will have a "hot spot" in angular position space.

To compare model parameters with curvature tuning, we considered two main cases based on the criterion of whether there was one dominant subunit or many. If the second largest weight was <70% of the largest weight, we considered the strongest subunit only (Fig. 10*C*). Otherwise, we considered the largest two subunits (Fig. 10*D*). We further divided the curvature tuning comparison into two cases based on the criterion of whether the absolute value of tuned curvature was higher or lower than 0.7 (as defined by the curvature scale in Pasupathy and Connor 2001). Because curvature is defined as a change in tangential angle over arc length, we computed the joint distributions of the differences in subunit orientations (roughly corresponding to the change in tangential angle) and the differences in angular positions of two subunits (roughly proportional to the arc length). There were only four discrete orientations for the C1 units in the model, and the orientation differences were binned by 0, 90, and 45/135° (the differences of 45 and 135° are ill defined). The angular position differences were binned by small, medium, and large differences (indicated by S, M, and L in the label) in 60° steps.

Figure 10, *C* and *D,* shows that some curvature tuning can be characterized by simple geometric relationships between C1 afferents. When there is one dominant subunit, its orientation has a strong influence on whether the neuron is tuned for sharp

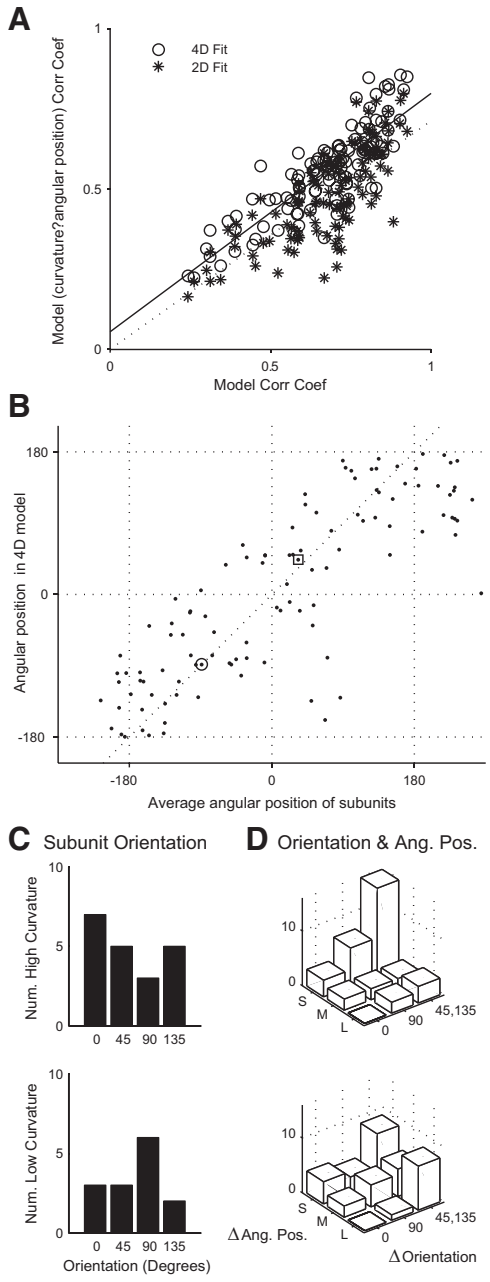

FIG. 10. Comparison of the C2 model and the boundary conformation model. *A*: comparison of goodness of fits of our V4 model and two boundary conformation tuning models (2D and 4D curvature and angular-position models) described in Pasupathy and Connor (2001). *B*: angular position of the boundary conformation tuning function correlates with the "center of mass" of all subunits (weighted by synaptic weight). The example neurons of Figs. 3 and 4 are indicated by □ and ○, respectively (at 45 and −90°). *C* and *D*: comparison of C1 subunit parameters with curvature tuning. Neurons are separated for this analysis into those that are tuned to high curvature (*top*) and low curvature (*bottom*) values and models with 1 dominant subunit (*1st column*) and many dominant subunits (*2nd column).* High curvature tuning can be achieved by a single dominant subunit oriented approximately radially, as seen in *C top* (the subunit orientation with respect to its angular position is 0°, similar to the example C2 unit in Fig. 3). If the subunit orientation was at 90° with respect to its angular position, the C2 unit tends to be tuned to the low curvature values, as shown in *C bottom*. When there are multiple dominant subunits, the two strongest subunits are considered for simplicity in *D*. The joint distributions for the difference in subunit orientations and angular positions (binned into small, S, medium, M, and large, L, differences) are shown. Low curvature tuning tends to arise from a large angular position separation (arc length) between the subunits, as indicated by the skewed (toward larger angular position differences) joint histogram in *D bottom* (an example is the C2 unit in Fig. 4). The results indicate that although we can identify some trends of correspondence between these two different models, the correspondence is not always straightforward.

or broad curvature fragments. If the subunit orientation and its angular position are parallel (for example, see Fig. 3*C*), the neuron generally produces high responses to sharp curvature fragments, which is evident from the bias toward 0° in Fig. 10*C, top*. If they are orthogonal, then the neuron is generally tuned for low curvature values, which is evident from the bias toward 90° in Fig. 10*C, bottom*. When multiple subunits have strong weights (like the example neuron in Fig. 4), the differences in their orientations and angular positions affect the curvature tuning, since curvature is determined by the rate of change in the tangent angle over the arc length. For the low curvature-tuned neurons, the two strongest subunits tend to have different orientations, and the angular position differences (proportional to the arc length) tend to be large (Fig. 10*D, top*).

Note that this analysis also shows that the correspondence between these two models is not always straightforward. For example, some neurons that exhibit tuning to high curvature and are fit with C2 units with one dominant C1 unit, have subunit orientations that are perpendicular to the radial direction instead of parallel. A full description of a C2 unit's tuning properties requires the inclusion of all the C1 afferents, and the approximations we have used here may not capture the full situation. Nonetheless, the geometric arrangement of oriented V1-like afferents (C1 units) can explain the observed curvature and angular position tuning behavior in many V4 neurons.

DISCUSSION

Our simulations demonstrate that a quantitative model of the ventral stream, theoretically motivated and biologically plausible, reproduces visual shape selectivity and invariance properties of area V4 from the known properties of lower visual area V1. The model achieves V4-like representations through a nonlinear, translation-invariant combination of locally selective subunits, suggesting a computational mechanism within or culminating in area V4. The simulated C2 units successfully reproduce selectivity and invariance to local translation for 109 V4 neurons tested with boundary conformation stimuli. Across the neural population, the model produces an average test set correlation coefficient of 0.57 (uncorrected for explainable variance). We also found that the population of C2 units qualitatively generalizes to other experimental stimulus sets using bars and complex gratings.

C2 units may form an intermediate code for representing boundary conformations in natural images. Figure 11 shows the responses of the two C2 units presented in Figs. 3 and 4 to two natural images. Based on the observed tuning properties of these neurons, it is not surprising to see that the first C2 unit responds strongly to the upper fins in the dolphin images, which contain sharp convex projections toward the upper right direction. The second C2 unit, which is selective for concave fragments in the lower portion of its receptive field, yields strong responses to several such boundary elements within the dolphin images. The graded responses of C2 unit populations may then form a representation of natural images that is particularly tuned to the conformations of various contours within an image. This code may be equivalent to the description provided by a previous study that demonstrated how a population code of V4 tuning functions could effectively represent contour stimuli (Pasupathy and Connor 2002). As seen in the two example images here, C2 responses can represent complex shapes or objects, even when curves and edges are difficult to define or segment and when the informative features are embedded within the boundary of an object (e.g., eyes, mouth, and nose within a face). Demonstrating this point, C2 units have been used as visual features to perform robust object recognition in natural images (Serre et al. 2007a,b). These results may suggest that V4 model neurons can respond like, and therefore be considered as, boundary conformation filters just as V1 neurons can be considered edge or orientation filters (Chisum and Fitzpatrick 2004; Daugman 1980; Jones and Palmer 1987; Mahon and De Valois 2001; Ringach 2004).

Our model of V4 is congruent with the major findings in Gallant et al. (1996)'s study, which indicate a bias within the population of V4 neurons to non-Cartesian gratings. Gallant et al. (1996) also proposed a mechanism, analogous to the simple to complex cell transformation in V1 proposed by Hubel and Wiesel (1962) to account for V4 responses. The ability of our model to predict the responses of a novel stimulus class given the responses of a training stimulus set suggests critical future tests for the model: test the grating selectivity predictions of C2 units, which were derived from V4 measurements using boundary conformation stimuli, against the physiologically measured responses of these same V4 neurons to gratings. In addition, our model of V4 can be shown (Serre et al. 2005) to reproduce the experimental data of Reynolds et al. (1999) for the condition without attentional modulation in V4. Although our model is not designed to address some other known properties of V4 responses, namely spectral and color selectivity (Schein and Desimone 1990), three-dimensional orientation tuning (Hinkle and Connor 2002), saliency (Mazer and Gallant 2003), or attentional effects (Reynolds et al. 1999), it accounts for much of the structural object-dependent selectivity and invariance currently described.

A recent publication (David et al. 2006) proposed that V4 response properties could be described with a second-order nonlinearity, called the spectral receptive field (SRF). This

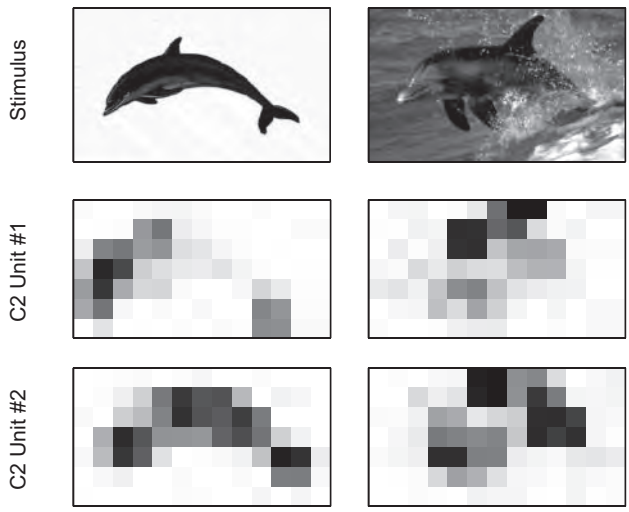

FIG. 11. Model responses to natural images. Two images of dolphins (*top*) and the responses of the two example C2 units from Figs. 3 (*middle*) and 4 (*bottom*) to these images are shown. Because C2 receptive fields cover a small fraction of these images, the response of a C2 unit was calculated on overlapping (by a factor of 1/3 the C2 receptive field) crops of the stimulus image. Based on their shape selectivity, these model C2 units respond strongly to certain features within the image, as indicated by the gray-scale response maps in the *middle* and *bottom* (dark areas indicate a high response). The images are from Fei-Fei et al. (2004).

description of V4 neurons is phenomenological and aimed at providing a robust regression model of the neural response, whereas our model is motivated and constrained by the computational goal of explaining object recognition in the ventral stream. It is therefore interesting to ask whether a connection exists between the two descriptions at the level of V4 cells. In fact, Volterra series analysis reveals that the leading term of our model is similar to the SRF (involving the spectral power of the input pattern), but the series associated with our model contains additional terms that are not negligible. In this sense, the model described here (Fig. 1) could be considered as similar but not identical to the model of David et al. (see APPENDIX). The additional aspects of our model describe some important aspects of V4 responses that are not described by the SRF. Because the SRF model lacks the spatial organization of afferent inputs, its response profiles will not be selective for angular position tuning, sensitive to the relative positions of features in space, or inhomogeneous within the receptive field, which are all attributes of C2 units. Our model architecture control also demonstrates the advantage of our two layer network for describing both selectivity and invariance in V4. Furthermore, the nonlinear selectivity operation (*Eqs. 1* and *2*) used by S2 units and the additional C2 layer account for the nonlinear summation properties of V4 (Desimone and Schein 1987; Gawne and Martin 2002; Gustavsen et al. 2004), which are not described by the SRF model. However, whereas our C2 model assumes a specific type of architecture and a set of nonlinear operations to explain the properties of the V4 neurons, the SRF model provides a more general and agnostic regression framework, which can be used to analyze and predict the neural responses not just specific to V4. The two models should ultimately be evaluated against experimental data. The correlation between predicted and actual data for the two models (0.32 for David et al. 2006 and 0.57 for our model) cannot be directly compared because the stimulus set used in David et al. (2006) is more complex and varied.

Learning may also play a critical role in the selectivity of V4 neurons. In our full model of the ventral pathway (see Fig. 1, *right*), the configurations and weights between S2 units and their oriented C1 afferents, which determine the selectivity of the C2 units, are learned from natural viewing experiences by a simple, unsupervised learning mechanism. According to our simulations, such learning mechanisms are capable of generating rich intermediate feature selectivities that account for the observed selectivity of V4 neurons (see section 4.2 of Serre et al. 2005; Serre et al. 2007a). Building on such intermediate feature selectivity, the model of the ventral pathway can perform object recognition tasks on natural images at performance levels at least as good as state-of-the-art image recognition algorithms and can mimic human performance in rapid categorization tasks (Serre et al. 2005, 2007a,b). The invariance may also be learned in a biophysically plausible way (e.g., Foldiak 1991; Wallis 1996; Wiskott and Sejnowski 2002), during a developmental period, from natural viewing experiences, such as watching a temporal sequence of moving objects. If temporally correlated neurons in a neighborhood connect to the same higher-order cell, the appropriate connectivity found between S2 and C2 units in the model can be generated (Serre et al. 2005).

Although our model is generally consistent with the known anatomical connectivity between V4 and lower visual areas, the full picture is certainly more complex. Beyond the description of a hierarchy of visual areas (Felleman and Van Essen 1991), the full anatomical picture includes "bypass" connections and highly organized inputs from V2. Connections from V1 to V4 that skip V2, known as bypass connections, represent a small but significant input to V4 (Nakamura et al. 1993). These connections may indicate two distinct inputs to V4 or may be considered as evidence for similar representations in V1 and V2 that are processed similarly in V4. In addition Shipp and Zeki (1995) and Xiao et al. (1999) have described the segregation and convergence of thin stripe and interstripe V2 regions onto V4. In light of these anatomical findings, it will be informative to determine if anatomically distinct inputs to V4 produce functionally distinct populations of neurons within V4. Overall, more work needs to be done to link the functional properties of V4 neurons and the anatomical connections between afferent areas.

How does V2 fit into our model of V4? There are relatively few experimental and theoretical studies of V2, making it difficult to include concrete constraints in our analysis. However, three hypotheses about the roles and functions of V2 are suggested by our hierarchical model. First, the selectivity and invariance seen in V4 may be constructed from yet another intermediate representation in V2, which itself is both more selective and more invariant than V1 (Ito and Komatsu 2004; Mahon and De Valois 2001) but less selective and less invariant than V4 (producing a continuum of receptive field sizes and invariance ranges depending on pooling ranges within the model), or second, V2 neurons are analogous to S2 units of the model so that they have complex shape selectivity but weak translation invariance [note that there may also be hypercomplex selectivity properties already present in V1 as reported by Mahon and De Valois (2001) and Hegde and Van Essen (2006)]. The more invariant representation is then realized by V4 neurons pooling over V2 neurons. Under this hypothesis, the cortico-cortical projections between areas V2 and V4 would represent fundamentally different transformations from the projections between V1 and V2. Third, area V2 is representationally similar to V1 for feedforward responses. Under this last hypothesis, area V4 may contain neurons analogous to both S2 and C2 units in the model or the selectivity representations (of S2 units) are computed through dendritic computations within neurons in V4 (Mel et al. 1998; Zhang et al. 1993). Experimental findings show that the majority of measured V4 responses are invariant to local translation, supporting the hypotheses that S2-like selectivity representations with small invariance range are present in another area of the brain, that they are computed implicitly in V4, or that there has been an experimental sampling bias. However, although V2 neurons are known to show selectivity over a range of stimulus sets (Hegde and Van Essen 2003; Ito and Komatsu 2004), there is not enough experimental data so far to verify or even distinguish these hypotheses. Carefully measuring and comparing both selectivity and invariance of areas V2 and V4 would be necessary to resolve this issue.

The V4 dataset examined from Pasupathy and Connor (2001) contained recordings using only one stimulus class and did not allow us to test the generalization abilities of the model to other types of stimuli. Although attempts were made to gauge the generalization capacity of the model (using cross-validation within the boundary conformation stimulus set and observing model responses to gratings and natural images), the ultimate

validation will require testing across a wider range of stimulus sets, including natural images. Furthermore, the current model is applicable only to the response of V4 neurons due to feedforward inputs and does not explain attentional or top-down factors (Mazer and Gallant 2003; Reynolds et al. 1999).

Our analysis of the representations in V4 adds to the mounting evidence for canonical circuits present within the visual system. Interestingly, our proposed mechanism for selectivity in V4 (a normalized weighted summation over the inputs, *Eq. 1*) is quite similar to the model of MT cells proposed in a recent publication (Rust et al. 2006). In addition, another recent study claims that motion integration in MT requires a local mechanism (Majaj et al. 2007), which may be analogous to our locally selective S2 units and more "global" C2 units for describing V4. Consequently, the same tuning and invariance operations may also be operating along the dorsal stream and may have a key role in determining various properties of motion-selective neurons in MT. Our model of V4 is also consistent with widely held beliefs on the ventral pathway, where more complex selectivity and a greater range of invariance properties are thought to be generated by precise combinations of afferent inputs. Previous quantitative studies have argued for similar mechanisms in other parts of the ventral stream (Perrett and Oram 1993). Further experimental work using parameterized shape spaces has shown that IT responses can be explained as a combination of invariant V4-like representations (Brincat and Connor 2004), which is consistent with our model (Serre et al. 2005). It has also been suggested that a tuning operation, used repeatedly in our model, may be a suitable mechanism for producing generalization, a key attribute of any learning system (Poggio and Bizzi 2004). Therefore instead of a collection of unrelated areas performing distinct tasks, the ventral pathway may be a system organized around two basic computational mechanisms necessary for robust object recognition.

APPENDIX

Comparison with the SRF model of V4

A recent publication (David et al. 2006) presented a general regression model on a very large set of neural responses and demonstrated that V4 response properties could be described in terms of a second-order nonlinear model, called the SRF. In this SRF framework, a V4 cell's response is analyzed by linearly combining the frequency components of the spatial autocorrelation of its inputs

$$r = \int \tilde{h}(\omega)\,|S(\omega)|^2 \mathrm{d}\mu(\omega) + r_0 + \varepsilon$$

where r is the response of the V4 neuron, $|S(\omega)|^2$ is the Fourier power spectrum of the visual pattern used as stimulus, ω is the two-dimensional vector of spatial frequencies. The SRF $\tilde{h}(\omega)$ is estimated from the data. The model of David et al. is closely related to energy models (Adelson and Bergen 1985; see also Poggio and Reichardt 1973). The underlying assumptions are as follows. The output of a simple cell in V1 centered in x is represented as a convolution of the stimulus s with a linear receptive field structure $h(x)$

$$\int h(x-\xi)s(\xi)\mathrm{d}\mu(\xi)$$

A complex cell in V1 is then described as the sum of the square of the output of simple cells of the same orientations at different positions within a neighborhood N (which introduces phase-invariance)

$$\sum_{S\in N}(h * s)(h * s) = \int \mathrm{d}\mu(\omega)\delta(\omega) \int \mathrm{d}\mu(\omega_1)S(\omega_1)H(\omega_1)S(\omega-\omega_1)H(\omega-\omega_1)$$

In our model, the simple cells are also described with a linear, orientation-selective convolution operation, plus a rectifying nonlinearity applied to ON-OFF and OFF-ON cells. Effectively, we perform an absolute value operation (invariant to contrast reversal) on the output of linear filtering

$$\left| \int h(x-\xi)s(\xi)\mathrm{d}\mu(\xi) \right|$$

A C1 unit, corresponding to a complex V1 cell, performs a max operation on the rectified output of a set of simple cells of the same orientations at different positions (and scales) within a neighborhood N.

From here on,[2] we use a rather general representation of nonlinear systems, the Volterra series (see Bedrosian and Rice 1971; Wu et al. 2006); for an analysis of its range of validity, see Palm and Poggio 1977). The Volterra series is a functional power series expansion containing linear and in general an infinite number of higher order terms. Although the multi-input version of the Volterra series should be used here, one may still assume the same one-input spatial frequency description of David et al. In this case, we can use Eq. 10 in (Bedrosian and Rice 1971)

$$Y(\omega) = G_1(\omega)S(\omega) + \frac{1}{2!}\int d\mu(\omega_1)G_2(\omega_1,\omega-\omega_1)S(\omega_1)S(\omega-\omega_1) + \frac{1}{3!}\int\int \times d\mu(\omega_1)d\mu(\omega_2)G_3(\omega_1,\omega_2,\omega-\omega_1-\omega_2)S(\omega_1)S(\omega_2)S(\omega-\omega_1-\omega_2) + \ldots$$

to describe the output of a C1 unit in our model. The max operation of a C1 unit can be approximated as the power series expansion of the softmax (Serre et al. 2005) or even more crudely as an average operation. Hence, the output of a C1 unit can be described as $Y(0)$ in the Fourier transform of the Volterra series above. Because of the absolute value operation, the series consists of even order terms only, and the response of a specific complex cell k is given by (with constant multiplicative factors omitted)

$$c_k = \int \mathrm{d}\mu(\omega_1)G_2^k(\omega_1,-\omega_1)|S(\omega_1)|^2 + \int\int \mathrm{d}\mu(\omega_1)\mathrm{d}\mu(\omega_2)|S(\omega_1)|^2|S(\omega_2)|^2G_4^k(\omega_1,-\omega_1,\omega_2,-\omega_2) + \ldots$$

A S2 unit in Fig. 1 combines the output of several C1 units $c_1, c_2, \ldots, c_M$ with a normalized dot product, yielding

$$r = \sum_{k=1,\ldots,M}\alpha_k c_k = \sum_{k=1,\ldots,M}\alpha_k \left(\int \mathrm{d}\mu(\omega_1)G_2^k(\omega_1,-\omega_1)\left|S(\omega_1)\right|^2 + \int\int \mathrm{d}\mu(\omega_1)\mathrm{d}\mu(\omega_2)\left|S(\omega_1)\right|^2\left|S(\omega_2)\right|^2 G_4^k(\omega_1,-\omega_1,\omega_2,-\omega_2) + \ldots \right)$$

where terms of degree >2 in general are not negligible. We expect a similar expansion to hold for C2 units because they inherit the tuning properties of S2 units.

[2] Although the Volterra series may be used from the earlier stages in the model, applying the Volterra expansion at the C1 level simplifies the analysis and allows an easier comparison between the models of David et al. and ours.

We see that the David et al. model corresponds to assuming that all kernels G are identically zero apart from $\sum_{k=1,\ldots,M} G_2^k$ and that the latter has the special form

$$G_2(\omega_1,\omega_2) = H(\omega_1)H(\omega_2)$$

This corresponds to linear filtering followed by a squaring operation. It is interesting that the leading term in both our and David et al. model is similar, involving the spectrum of the input pattern (although in the case of our model the 2nd-order term does not necessarily have the simple form of David et al.). The additional terms in the equation above are specifically dictated by our model architecture; the only parameters we estimate in this paper are the number M and type of subunits for each C2 unit from a fixed set.

ACKNOWLEDGMENTS

The authors thank D. Zoccolan, G. Kreiman, T. Serre, and U. Knoblich for valuable discussions regarding this work.

GRANTS

This research was sponsored by grants from DARPA, National Institute of Mental Health Grant P20MH-66239, Office of Naval Research and the National Science Foundation. Additional support was provided by Honda Research Institute, Sony, Gerry Burnett, and the McDermott chair (T. Poggio). C. Connor was sponsored by the Pew Scholars Program in the Biomedical Sciences and National Institutes of Health (NINDS and NEI).

REFERENCES

Adelson EH, Bergen JR. Spatiotemporal energy models for the perception of motion. *J Opt Soc Am A Opt Image Sci Vision* 2: 284–299, 1985.

Baizer JS, Robinson DL, Dow BM. Visual responses of area 18 neurons in awake, behaving monkey. *J Neurophysiol* 40: 1024–1037, 1977.

Bedrosian E, Rice SO. The output properties of Volterra systems (nonlinear systems with memory) driven by harmonic and Gaussian inputs. *Proc IEEE* 59: 1688–1707, 1971.

Brincat SL, Connor CE. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nat Neurosci* 7: 880–886, 2004.

Brincat SL, Connor CE. Dynamic shape synthesis in posterior inferotemporal cortex. *Neuron* 49: 17–24, 2006.

Burges CJC. A tutorial on Support Vector Machines for pattern recognition. *Data Mining Knowl Disc* 2: 121–167, 1998.

Cadieu C, Kouh M, Riesenhuber M, and Poggio T. *Shape Representation in V4: Investigating Position-Specific Tuning for Boundary Conformation with the Standard Model of Object Recognition.* CBCL Paper 241/AI Memo 2004–024. Cambridge, MA: MIT, 2004.

Carandini M, Heeger DJ, Movshon JA. Linearity and normalization in simple cells of the macaque primary visual cortex. *J Neurosci* 17: 8621–8644, 1997.

Chisum HJ, Fitzpatrick D. The contribution of vertical and horizontal connections to the receptive field center and surround in V1. *Neural Netw* 17: 681–693, 2004.

Daugman JG. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Res* 20: 847–856, 1980.

David SV, Hayden BY, Gallant J. Spectral receptive field properties explain shape selectivity in area V4. *J Neurophysiol* 96: 3492–3505, 2006.

De Valois RL, Yund EW, Hepler N. The orientation and direction selectivity of cells in macaque visual cortex. *Vision Res* 22: 531–544, 1982.

De Weerd P, Desimone R, Ungerleider LG. Cue-dependent deficits in grating orientation discrimination after V4 lesions in macaques. *Vis Neurosci* 13: 529–538, 1996.

Desimone R, Schein SJ. Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. *J Neurophysiol* 57: 835–868, 1987.

Fei-Fei L, Fergus R, Perona P. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: *IEEE CVPR 2004, Workshop on Generative-Model Based Vision* 2004, p. 178.

Felleman DJ, Van Essen DC. Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex* 1: 1–47, 1991.

Foldiak P. Learning invariance from transformation sequences. *Neural Comput* 3: 194–200, 1991.

Freiwald WA, Tsao DY, Tootell RBH, Livingstone MS. Complex and dynamic receptive field structure in macaque cortical area V4d. *J Vision* 4: 184–184, 2004.

Fukushima K, Miyake S, Ito T. Neocognitron: a neural network model for a mechanism of visual pattern recognition. *IEEE Trans Syst Man Cybern* 13: 826–834, 1983.

Gallant JL, Connor CE, Rakshit S, Lewis JW, Van Essen DC. Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *J Neurophysiol* 76: 2718–2739, 1996.

Gallant JL, Shoup RE, Mazer JA. A human extrastriate area functionally homologous to macaque V4. *Neuron* 27: 227–235, 2000.

Gawne TJ, Martin JM. Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. *J Neurophysiol* 88: 1128–1135, 2002.

Girard P, Lomber SG, Bullier J. Shape discrimination deficits during reversible deactivation of area V4 in the macaque monkey. *Cereb Cortex* 12: 1146–1156, 2002.

Gross CG, Rocha-Miranda CE, Bender DB. Visual properties of neurons in inferotemporal cortex of the macaque. *J Neurophysiol* 35: 96–111, 1972.

Gustavsen K, David SV, Mazer JA, and Gallant J. Stimulus interactions in V4: a comparison of linear, quadratic, and max models. *Soc Neurosci Abstr* 2004.

Heeger DJ. Modeling simple-cell direction selectivity with normalized, half-squared, linear operators. *J Neurophysiol* 70: 1885–1898, 1993.

Hegde J, Van Essen DC. Strategies of shape representation in macaque visual area V2. *Vis Neurosci* 20: 313–328, 2003.

Hegde J, Van Essen DC. A comparative study of shape representation in Macaque visual areas V2 and V4. *Cereb Cortex* 17: 1100–1116, 2006.

Hinkle DA, Connor CE. Three-dimensional orientation tuning in macaque area V4. *Nat Neurosci* 5: 665–670, 2002.

Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* 160: 106–154, 1962.

Hubel DH, Wiesel TN. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J Neurophysiol* 28: 229–289, 1965.

Hubel DH, Wiesel TN. Receptive fields and functional architecture of monkey striate cortex. *J Physiol* 195: 215–243, 1968.

Hung CP, Kreiman G, Poggio T, DiCarlo JJ. Fast readout of object identity from macaque inferior temporal cortex. *Science* 310: 863–866, 2005.

Ito M, Komatsu H. Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *J Neurosci* 24: 3313–3324, 2004.

Jones JP, Palmer LA. The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *J Neurophysiol* 58: 1187–1211, 1987.

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.

Kouh M, Riesenhuber M. Investigating shape representation in area V4 with HMAX: orientation and grating selectivities. *CBCL Paper 231/AI Memo 2003–021.* Cambridge, MA: MIT, 2003.

Li Z. A neural model of contour integration in the primary visual cortex. *Neural Comput* 10: 903–940, 1998.

Logothetis NK, Pauls J, Poggio T. Shape representation in the inferior temporal cortex of monkeys. *Curr Biol* 5: 552–563, 1995.

Mahon LE, De Valois RL. Cartesian and non-Cartesian responses in LGN, V1, and V2 cells. *Vis Neurosci* 18: 973–981, 2001.

Majaj NJ, Carandini M, Movshon JA. Motion integration by neurons in Macaque MT is local, not global. *J Neurosci* 27: 366–370, 2007.

Mazer JA, Gallant JL. Goal-related activity in V4 during free viewing visual search. Evidence for a ventral stream visual salience map. *Neuron* 40: 1241–1250, 2003.

Mel BW. SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comput* 9: 777–804, 1997.

Mel BW, Fiser J. Minimizing binding errors using learned conjunctive features. *Neural Comput* 12: 247–278, 2000.

Mel BW, Ruderman DL, Archie KA. Translation-invariant orientation tuning in visual "complex" cells could derive from intradendritic computations. *J Neurosci* 18: 4325–4334, 1998.

Merigan WH, Pham HA. V4 lesions in macaques affect both single- and multiple-viewpoint shape discriminations. *Vis Neurosci* 15: 359–367, 1998.

Nakamura H, Gattass R, Desimone R, Ungerleider LG. The modular organization of projections from areas V1 and V2 to areas V4 and TEO in macaques. *J Neurosci* 13: 3681–3691, 1993.

Palm G, Poggio T. Volterra representation and Wiener expansion—validity and pitfalls. *Siam J Appl Math* 33: 195–216, 1977.

Pasupathy A, Connor CE. Responses to contour features in macaque area V4. *J Neurophysiol* 82: 2490–2502, 1999.

Pasupathy A, Connor CE. Shape representation in area V4: position-specific tuning for boundary conformation. *J Neurophysiol* 86: 2505–2519, 2001.

Pasupathy A, Connor CE. Population coding of shape in area V4. *Nat Neurosci* 5: 1332–1338, 2002.

Perrett DI, Oram MW. Neurophysiology of shape processing. *Image Vision Comput* 11: 317–333, 1993.

Poggio T, Bizzi E. Generalization in vision and motor control. *Nature* 431: 768–774, 2004.

Poggio T, Reichardt W. Considerations on models of movement detection. *Kybernetik* 13: 223–227, 1973.

Pollen DA, Przybyszewski AW, Rubin MA, Foote W. Spatial receptive field organization of macaque V4 neurons. *Cereb Cortex* 12: 601–616, 2002.

Reynolds JH, Chelazzi L, Desimone R. Competitive mechanisms subserve attention in macaque areas V2 and V4. *J Neurosci* 19: 1736–1753, 1999.

Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. *Nat Neurosci* 2: 1019–1025, 1999.

Ringach DL. Mapping receptive fields in primary visual cortex. *J Physiol* 558: 717–728, 2004.

Russell SJ, Norvig P. *Artifical Intelligence: A Modern Approach* (2nd ed.). New York: Prentice Hall, 2003.

Rust NC, Mante V, Simoncelli EP, Movshon JA. How MT cells analyze the motion of visual patterns. *Nat Neurosci* 9: 1421–1431, 2006.

Schein SJ, Desimone R. Spectral properties of V4 neurons in the macaque. *J Neurosci* 10: 3369–3389, 1990.

Schiller PH. Effect of lesions in visual cortical area V4 on the recognition of transformed objects. *Nature* 376: 342–344, 1995.

Schiller PH, Finlay BL, Volman SF. Quantitative studies of single-cell properties in monkey striate cortex. III. Spatial frequency. *J Neurophysiol* 39: 1334–1351, 1976.

Schiller PH, Lee K. The role of the primate extrastriate area V4 in vision. *Science* 251: 1251–1253, 1991.

Serre T, Kouh M, Cadieu C, Knoblich U, Kreiman G, and Poggio T. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. *CBCL Paper 259/AI Memo 2005–036.* Cambridge, MA: MIT, 2005.

Serre T, Oliva A, Poggio T. A feedforward architecture accounts for a rapid categorization. *Proc Natl Acad Sci USA* 104: 6424–6429, 2007a.

Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T. Robust object recognition with cortex-like mechanisms. *IEEE Trans Pattern Anal Mach Intell* 29: 411–426, 2007b.

Shipp S, Zeki S. Segregation and convergence of specialised pathways in macaque monkey visual cortex. *J Anatomy* 187: 547–562, 1995.

Tanaka K, Saito H, Fukada Y, Moriya M. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J Neurophysiol* 66: 170–189, 1991.

Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system. *Nature* 381: 520–522, 1996.

Ungerleider LG, Haxby JV. "What" and "where" in the human brain. *Curr Opin Neurobiol* 4: 157–165, 1994.

Wallis G. Using spatio-temporal correlations to learn invariant object recognition. *Neural Netw* 9: 1513–1519, 1996.

Wilson HR, Wilkinson F. Detection of global structure in glass patterns: implications for form vision. *Vision Res* 38: 2933–2947, 1998.

Wiskott L, Sejnowski TJ. Slow feature analysis: unsupervised learning of invariances. *Neural Comput* 14: 715–770, 2002.

Wu MC, David SV, Gallant JL. Complete functional characterization of sensory neurons by system identification. *Annu Rev Neurosci* 29: 477–505, 2006.

Xiao Y, Zych A, Felleman DJ. Segregation and convergence of functionally defined V2 thin stripe and interstripe compartment projections to area V4 of macaques. *Cereb Cortex* 9: 792–804, 1999.

Yu AJ, Giese MA, Poggio TA. Biophysiologically plausible implementations of the maximum operation. *Neural Comput* 14: 2857–2881, 2002.

Zhang KC, Sereno MI, Sereno ME. Emergence of position-independent detectors of sense of rotation and dilation with Hebbian learning—an analysis. *Neural Comput* 5: 597–612, 1993.