On the Road to Invariant Recognition: Explaining Tradeoff and Morph Properties of Cells in Inferotemporal Cortex using Multiple-Scale Task-Sensitive Attentive Learning

Stephen Grossberg, Jeffrey Markowitz, and Yongqiang Cao

Center for Adaptive Systems
Department of Cognitive and Neural Systems
Center of Excellence for Learning in Education, Science and Technology
Boston University
677 Beacon Street, Boston, MA, 02215, USA

Running Title: IT Learning, Recognition, Tradeoff, and Morph Properties

December 17, 2010
Revised: March 29, 2011

Neural Networks, in press

All correspondence should be addressed to
Professor Stephen Grossberg
Center for Adaptive Systems
Department of Cognitive and Neural Systems
Boston University
677 Beacon Street
Boston, MA 02215
Email: steve@bu.edu
Phone: 617-353-7858/7
Fax: 617-353-7755

Abstract

Visual object recognition is an essential accomplishment of advanced brains. Object recognition needs to be tolerant, or invariant, with respect to changes in object position, size, and view. In monkeys and humans, a key area for recognition is the anterior inferotemporal cortex (ITa). Recent neurophysiological data show that ITa cells with high object selectivity often have low position tolerance. We propose a neural model whose cells learn to simulate this tradeoff, as well as ITa responses to image morphs, while explaining how invariant recognition properties may arise in stages due to processes across multiple cortical areas. These processes include the cortical magnification factor, multiple receptive field sizes, and top-down attentive matching and learning properties that may be tuned by task requirements to attend to either concrete or abstract visual features with different levels of vigilance. The model predicts that data from the tradeoff and image morph tasks emerge from different levels of vigilance in the animals performing them. This result illustrates how different vigilance requirements of a task may change the course of category learning, notably the critical features that are attended and incorporated into learned category prototypes. The model outlines a path for developing an animal model of how defective vigilance control can lead to symptoms of various mental disorders, such as autism and amnesia.

Keywords: inferotemporal cortex; object recognition; categorization; attention; cortical magnification factor; multiple spatial scales; selectivity-tolerance tradeoff; image morphs; adaptive resonance theory; vigilance; autism; amnesia

Introduction

Humans and other primates effortlessly recognize objects in the world as they move their eyes, heads, and bodies with respect to them. This flexibility implies a high degree of invariance during object recognition. Multiple cortical areas, ranging from V1, V2, and V4 through inferotemporal and prefrontal cortex, gradually build up such an invariance in stages. One important stage occurs in the inferotemporal cortex (IT).

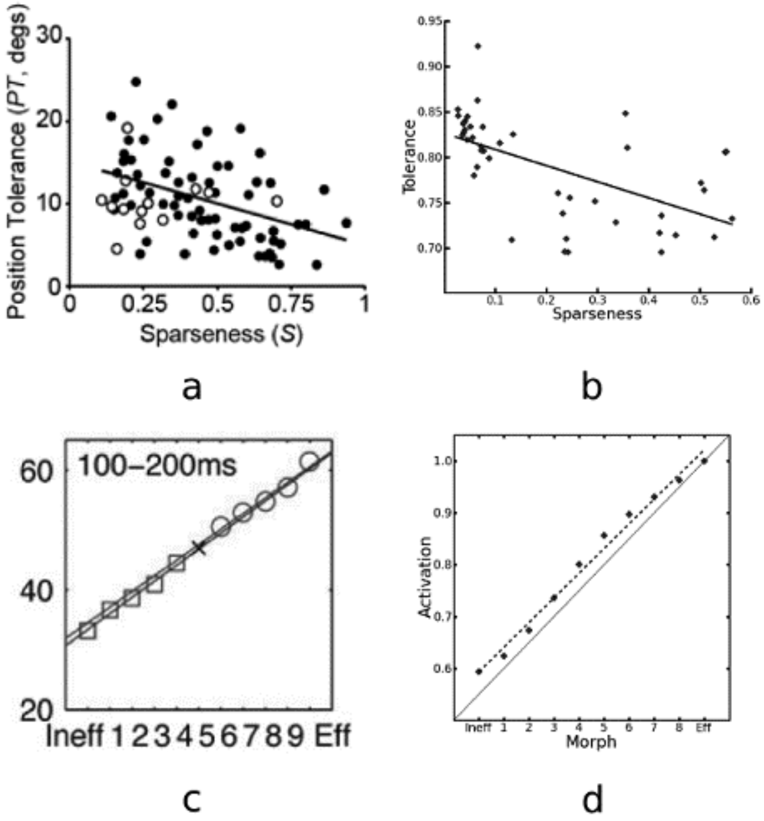

Figure 1. (a) Experimental data reprinted with permission from Zoccolan et al. (2007). (b) Response profiles of model ITa category cells learned at low vigilance (ρ = .1). (c) Experimental data reprinted with permission from Akrami et al. (2009), where the ordinate indicates spikes/sec. (d) Responses of model ITa category cells (dashed line) learned at high vigilance (ρ = .9).

Classical lesion studies have shown that IT supports visual object recognition in primates (Ettlinger, Iwai, Mishkin, & Rosvold, 1968). Single cell recordings show that some IT cells have large receptive fields (~23°) that contain the fovea and ipsi- and contra-lateral visual fields, and respond selectively to ‘complex’ objects; e.g., two-dimensional silhouettes of hands and Fourier descriptors (Schwartz, Desimone, Albright, & Gross, 1983; Gross, Miranda, & Bender, 1972). These properties are natural in a cortical area subserving invariant object recognition. However, recent experiments show that some IT cells have much smaller receptive fields (DiCarlo & Maunsell, 2003). Moreover, various cells in the anterior inferotemporal cortex (ITa), which has been thought to support invariant recognition, exhibit a *tradeoff* between *object selectivity* (or *sparseness*) and *position tolerance* (or *invariance*); namely, neurons with high object selectivity typically have low position tolerance and vice versa (Figure 1a; Zoccolan, Kouh, Poggio, & DiCarlo, 2007). How can such a tradeoff be reconciled with invariant object recognition?

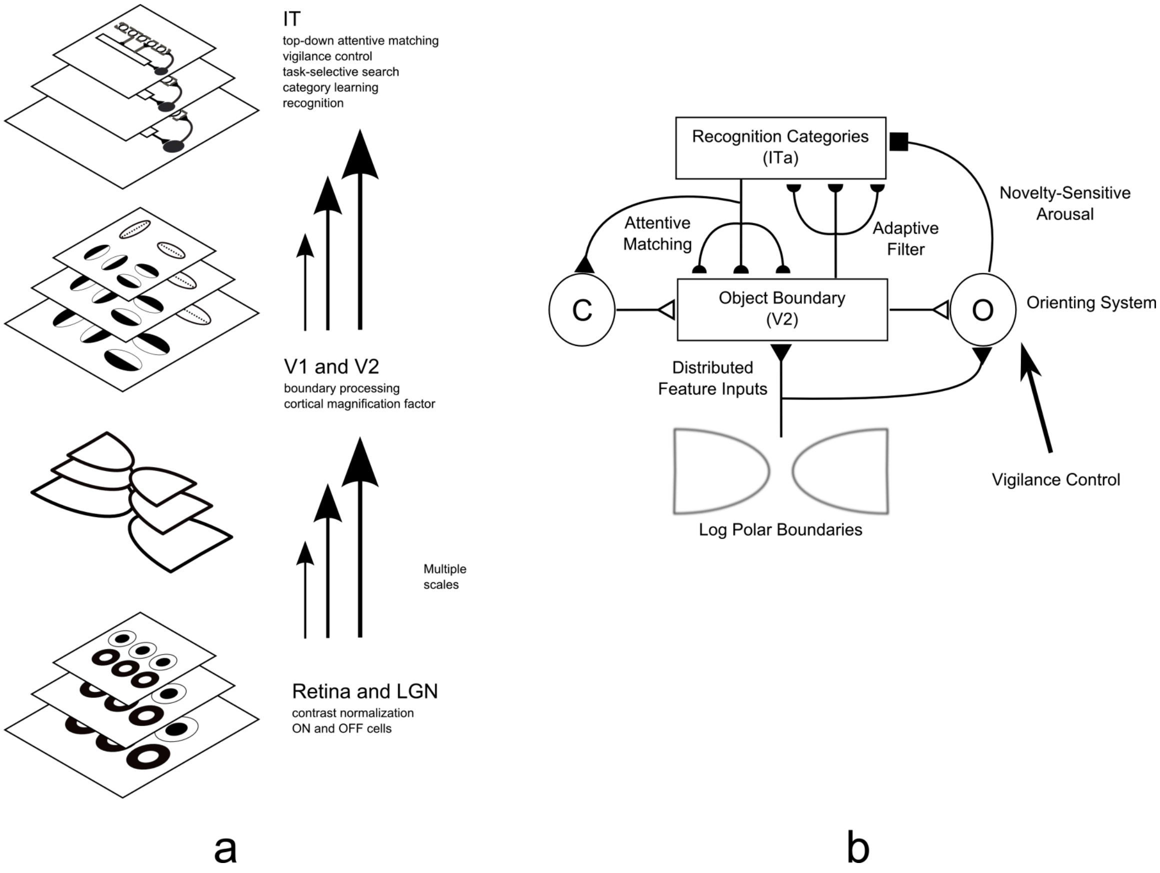

Figure 2. (a) Model circuit: LGN ON (OFF) cells map to cortex via the log-polar transform and are then boundary processed with simple (half filled) and complex (not filled) cells. Then, an attentive ART category learning and recognition network models responses of ITa cells. (b) Category learning circuit: Top-down expectations are attentively matched against bottom-up input features. Attended critical features are learned. A big enough mismatch (i.e., one that does not satisfy vigilance in the orienting system O) causes a burst of novelty-sensitive arousal from O that inhibits the active category and triggers search for and learning of a better matching category. C: biased competition; open triangles: inhibitory connections; filled triangles: excitatory connections; filled semi-circles: learned connections.

We propose a neural model (Figure 2) that explains how interactions within and between a series of processing stages in the ventral What cortical processing stream can produce this tradeoff. As in previous models (e.g., Fukushima, 1988; Watson, 1987; Zoccolan et al., 2007), visual inputs are preprocessed by perceptual mechanisms before the preprocessed representations activate recognition processes. In the current model, the perceptual preprocessing stages include contrast normalization of the input image by ON and OFF cells in the model retina and lateral geniculate nucleus (LGN). This contrast-normalized representation is then transformed into log-polar coordinates within the model visual cortex, wherein a boundary representation is generated by cortical simple and complex cells. These contrast-normalized, log-polar boundary signals then activate a model IT network whose interacting bottom-up adaptive filtering and top-down learned expectations can learn recognition categories under both unsupervised and supervised conditions. Interactions among the processes within this hierarchy of stages automatically generates the experimentally observed tradeoff as an emergent property.

The models described in Zoccolan et a. (2007) simulate the tradeoff in two different ways: (1) by varying the number of model V4 cells that project to individual IT cells, or (2) by varying the width of the multi-dimensional Gaussian tuning function used to model their IT cells**.** More precisely, parameters in the lower levels in their model hierarchy (i.e., V1, V2, and V4) were fixed, and model IT cells were manually tuned to maximally respond to particular objects in the image set. Introducing variability post-hoc in the tuning function or in the number of afferents to IT cells both produced the tradeoff. Our model shows, in contrast, how the tradeoff can emerge from the adaptive dynamics of system interactions without any manual selection of parameters.

Our model also simulates the data of Akrami et al. (2009) showing the responses of IT neurons to image morphs (Figure 1c).

There is a relationship between the Zoccolan et al. (2007) and Akrami et al. (2009) data sets that the model is used to clarify. This relationship depends crucially upon the fact that category learning and recognition in the model combine bottom-up learning and activation of recognition categories with category-activated read-out of top-down learned expectations that are matched against the object boundary representations (Figure 2). This matching process focuses attention on the subset of object features that are shared by the bottom-up input pattern and the top-down expectation. Such top-down matching, leading to a focus of object attention, is well-known to occur in IT from multiple behavioral, anatomical, and neurophysiological experiments (e.g., Desimone, 1998; Li, Miller, & Desimone, 1993). In the model, this top-down process can be modulated by a task-selective criterion of matching that is called *vigilance.* Different levels of vigilance determine whether the criterion for a good enough top-down attentive match is coarse (low vigilance) or fine (high vigilance). High vigilance enables the learning of concrete and specific recognition categories, whereas low vigilance enables the learning of abstract and general recognition categories (Carpenter & Grossberg, 1991, 1993; Grossberg, 2003, 2007). Using variable vigilance, the brain can learn recognition categories that match variable task demands.

Our model predicts that the task in the Zoccolan et al. (2007) experiment led to learning under low vigilance, since the monkeys in this task passively experienced their stimuli. In contrast, the Akrami et al. (2009) task led to learning under high vigilance, since these monkeys did active discrimination during a delayed match-to-sample task.

These two sets of results are thus proposed to demonstrate how task demands can influence the way in which attention may be allocated in different ways to support the learning of invariant recognition categories.

Thus, our model clarifies the role of top-down attention in the learning of recognition categories, and how task requirements can alter the criterion for a good enough match and thus the recognition categories that will be learned. Such computations are not realized by purely feedforward models, such as the model of Riesenhuber & Poggio (1999, 2002). Our results thus contribute to the general theme of how attentional task demands can alter cortical learned receptive field properties, and illustrate the importance of considering task-dependent factors when interpreting neurophysiological results in which animals carry out tasks of variable difficulty.

Model Overview

Model mechanisms and their functional rationale are listed intuitively below, along with links to the model equations that realize them mathematically in the Appendix. All of the model mechanisms have support from psychological and neurobiological data. The model's power derives from how quantitative properties that match challenging neurophysiological data may arise as emergent properties of all of these processes when they interact together. Readers can skip this section if they wish to first read the simulation results with the model diagram of Figure 2 in mind.

Image inputs. The model learns to categorize natural objects, chosen from the Cal Tech 101 image data base (Fei, Fergus, & Perona, 2006) that was used by Zoccolan et al. (2007); see Figure S1 in the Supplemental Information. Each image is processed by oversampled hemi-retinas (Figure 3). Oversampling includes part of the opposing hemi-retina near the vertical meridian, consistent with neurophysiological evidence (Van Essen, Newsome, & Maunsell, 1984), and prevents sampling artifacts near the vertical meridian (Fazl et al. (2009). Each position computes the average of the three RGB color values of the input image (see Appendix Equation (A1)).

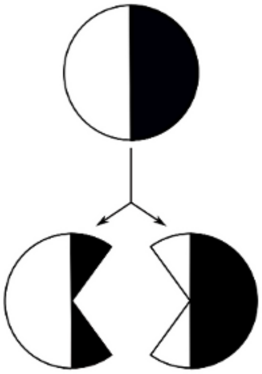

Figure 3. Schematic depiction of the hemi-retina split. The input image, a circular cut-out of the original image, is split into two hemi-retinas (black and white). In accordance

with experimental data, each hemi-retina fed to the model is oversampled by being augmented by part of the opposing hemi-retina near the vertical meridian.

ON and OFF cells and contrast normalization. Within the model's simplified retina and LGN, these inputs are processed by ON cells, which are turned on by image increments, and by OFF cells, which are turned off by image increments. These ON and OFF cells come in three receptive field sizes, from coarse to fine; see Appendix Equations (A2)-(A4). The existence of multiple scales is supported by both psychophysical and neurophysiological data (Foster, Gaska, Nagler, & Da Pollen, 1985; Julesz & Schumer, 1981; Movshon, Thompson, & Tolhurst, 1978; Tyler, 1975), and is shown below to play an important role in explaining the selectivity/tolerance tradeoff that was reported by Zoccolan et al. (2007). Multiple scales have been employed in a number of models of IT cortex and, more generally, the What cortical processing stream (Gochin, 1994; Riesenhuber & Poggio, 1999; Stringer, Perry, Rolls, & Proske, 2006; Serre et al., 2007). In our model, the ON and OFF cells obey the membrane, or shunting, equations of neurophysiology (Hodgkin, 1964) within on-center off-surround networks that are capable of contrast-normalizing cell responses to input images (Grossberg, 1973, 1980).

Log-polar cortical magnification. These contrast-normalized activity patterns at each spatial scale then undergo a log-polar transform (Appendix Equation (A5)) that computes the cortical magnification factor in the model's highly simplified cortical areas V1/V2 (Daniel & Whitteridge, 1961; Fischer, 1973; Horton & Hoyt, 1991; Schwartz, 1980; Tootell, Silverman, Switkes, & De Valois, 1982; Van Essen et al., 1984). The log-polar transform over-represents the fovea and under-represents the periphery, a critical point since Zoccolan et al. (2007) presented images up to 10° from the fovea. The log-polar transform operates on the signals from each hemi-retina, which are then concatenated into a single ON or OFF channel map; see Appendix Equations (A6) and (A7).

Simple and complex cells. The contrast-normalized and cortically magnified images in each scale input to simple cells in the model's cortical area V1 (Hubel & Wiesel, 1959). The simple cells act as filters that detect oriented features of a fixed contrast polarity in the image along the filter's preferred orientation. Each simple cell receptive field is activated by LGN ON and OFF outputs (Clay & Alonso, 1995; Hirsch, Alonso, Reid, & Martinez, 1998; Hubel & Wiesel, 1962) that are filtered by spatially elongated and offset Gaussian kernels. In particular, the ON and OFF inputs are converted into oriented simple cell responses using oriented, multiple-scale, difference-of-offset-Gaussian filters that respond to a particular contrast polarity within a self-normalizing recurrent shunting network; see Appendix Equations (A8)-(A10). In model simulations, simple cells at each position are selective to four different orientations at each of the three spatial scales. This simple cell model has previously been used in simulations of psychophysical and neurophysiological data (e.g., Bhatt et al., 2007; Grossberg & Raizada, 2000; Grossberg et al., 2007), as well as to process complex images (e.g., Mingolla et al., 1999). Then the output signals of pairs of simple cells that are sensitive to opposite contrast polarities at each position, orientation, and scale are combined to activate complex cells whose responses are contrast-invariant; see Appendix Equation (A11).

Object boundaries. To compute object boundaries in the model cortical area V2, the complex cell output signals at each position and scale are added across the four different orientations to derive a measure of unoriented boundary strength within each scale. To quantitatively fit the targeted neurophysiological data, no further boundary processing is needed.

In particular, depth-selective boundaries are not needed. Nor are processes of boundary completion. A great deal is now known about how depth-selective boundaries may be formed, completed, and used to separate figures from their backgrounds in cases where more ambiguous natural images may require additional stages of processing; e.g., Cao & Grossberg (2005, 2011); Fang & Grossberg (2009); Grossberg & Yazdanbaksh (2005); Kelly & Grossberg (2000).

Many psychophysical studies have supported the prediction that depth-selective boundaries and surfaces are the perceptual units of 3D vision (Grossberg, 1987). Boundaries are often sufficient to enable object recognition (Alvarez & Cavanagh, 2008; Davidoff, 1991; Elder & Zucker, 1998; Grossberg, 1994; Grossberg & Mingolla, 1985; von der Heydt & Peterhans, 1989; von der Heydt, Peterhans, & Baumgartner, 1984; Lamme, Rodriguez-Rodriguez, & Spekreijse, 1999; Rogers-Ramachandran & Ramachandran, 1998), and that is the case for the targeted data.

Control of category learning and recognition by top-down attentive matching, vigilance control, and memory search. Contrast-normalized, log-polar transformed boundaries are the inputs to the model IT, which carries out incremental, fast learning of recognition categories whose bottom-up filters and learned top-down expectations learn prototypes of attended critical features (Bhatt et al., 2007; Carpenter & Grossberg, 1991, 1993; Grossberg, 1980, 2007). The top-down expectations focus attention on the pattern of critical features that is learned by the prototype of the expectation (Figure 2). Attention carries out a top-down matching process (cf. Miller, Li, & Desimone, 1993), which computes how well the prototype matches a bottom-up feature pattern. This attentive matching process is realized by a top-down, modulatory on-center, driving off-surround network (Figure 2), which explains, and indeed predicted, data properties of self-normalizing "biased competition" (Desimone, 1998; Reynolds & Heeger, 2009).

The interactions of these bottom-up and top-down processes helps to solve the *stability-plasticity dilemma* (Grossberg, 1980); that is, to enable brains to learn quickly without forcing rapid and unselective, or "catastrophic", forgetting. If a match is good enough to cause a bottom-up/top-down resonance between the attended critical feature pattern and the selected category, then this resonance triggers learning in the adaptive weights, or long-term memory traces, that occur in the pathways that carry signals between the attended features and the selected category. Such a resonance embodies a system-wide consensus that the critical feature patterns are worthy of being learned, and dynamically buffers system memories against catastrophic forgetting. Hence the name Adaptive Resonance Theory, or ART, for the emerging theory that computationally characterizes how this may occur (e.g., Carpenter & Grossberg, 1991, 1993; Grossberg, 1980, 2003, 2007). If the match is not good enough, then a mismatch is computed between the learned top-down expectation and the bottom-up input feature pattern. A sufficiently large mismatch triggers a memory search, or bout of hypothesis testing, that leads to discovery and learning of a better-matching category.

The criterion for a good enough match can depend upon the task (cf. Spitzer, Desimone, & Moran, 1988). A task-selective *vigilance parameter* determines how strict the matching criterion is, with higher vigilance requiring a better match to trigger resonance and learning. (Carpenter & Grossberg, 1991, 1993; Grossberg & Merrill, 1996; Grossberg & Seidman, 2006). Hence high vigilance leads to the learning of more specific and concrete categories (e.g., a view category that codes similar poses of a single face), and low vigilance leads to the learning of more general and abstract categories (e.g., a face category that responds to multiple faces).

An *orienting system* computes the degree of match between the bottom-up input pattern and the learned top-down expectation (Figure 2). In other words, the orienting system calibrates the degree of novelty of the currently active input feature pattern relative to the currently active learned top-down expectation. If the current input is too novel to satisfy the vigilance criterion, then the orienting system sends a burst of nonspecific arousal to the category learning network, resets the currently active category and its top-down expectation, and frees the system to choose a different, and better-matching, category. For reviews of data supporting how the brain may use such top-down attentive matching, vigilance, and memory search mechanisms to learn cortical recognition codes, see Banquet & Grossberg (1987), Carpenter & Grossberg (1991, 1993), Engel et al. (2001), Grossberg (2003, 2007), Grossberg & Seidman (2006), Pollen (1999), and Raizada & Grossberg (2003).

In a detailed spiking laminar circuit model of thalamocortical interactions, Grossberg & Versace (2008) predicted how vigilance may be controlled by novelty-sensitive corticothalamic mismatches that activate the nonspecific thalamus (part of the orienting system), which in turn activates acetylcholine release via the nucleus basalis of Meynert, thereby increasing excitability of cortical layer 5 cells, and leading to cortical reset in layer 4 via layer 5-to-6-to-4 signals. Deficient vigilance control has been predicted to occur in autism and medial temporal amnesia, among other mental disorders (Carpenter & Grossberg, 1993; Grossberg & Seidman, 2006). If the above predictions about the anatomical and pharmacological substrates of vigilance control are confirmed, then manipulations in animals that mimic the predicted vigilance deficits may help to create an animal model of these disorders.

Results

Since the monkeys in the Zoccolan et al. experiment passively experienced their stimuli, vigilance was set low (ρ = .1) in simulations of their data. Since the monkeys in the Akrami et al. (2009) experiments did active discrimination during a delayed matching-to-sample task while experiencing image morphs, the vigilance was set high (ρ = .9) in simulations of these data. Figures 1a and 1c summarize the data from Zoccolan et al. (2007) and Akrami et al. (2009). Figures 1b and 1d summarize the model simulations of these data, respectively.

Zoccolan et al. (2007) data simulations. The stimuli used in these simulations were 203 Cal Tech 101 images from the experiment at, below, and above the center of gaze (see Figure S1 in the Supplemental Information). For each run, 200 random samples without replacement were used during which the category learning process discovered and learned new categories. During testing, all 203 objects were presented at the center to measure selectivity, and 49 objects were presented at all three training positions to measure position tolerance.

To measure a category cell's selectivity, we used the same metric that was used as in Zoccolan et al. (2007):

$$S = \left[1 - \frac{\left[\frac{\left(\sum R_i \right)^2}{n} \right] \Big/ \left[\frac{\sum R_i^2}{n} \right]}{1 - \frac{1}{n}} \right], (1)$$

where R_i is the response of a category to input *i* and *n* is the total number of stimuli (n = 203). To measure tolerance, the mean of the inverse of the standard deviation of a category cell's response was computed for all objects across the three positions. Because the standard deviation gives an estimate of response spread, its inverse provides a measure of tolerance, since the higher the spread the lower the tolerance, and vice-versa. Finally, the Pearson correlation coefficient was computed between the two values as in Zoccolan et al. (2007), thereby leading to the simulation result in Figure 1b.

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
Sim. 1	-.359	-.589	-.467	-.419	-.279	-.63	-.311	-.449	-.312	-.279
Sim. 2	-.397	-.570	-.518	-.573	-.401	-.196	-.245	-.332	-.332	-.338
Sim .3	-.558	-.350	-.333	-.415	-.430	-.333	-.688	-.478	-.468	-.487
Sim. 4	-.602	-.598	-.143	-.422	-.280	-.669	-.390	-.258	-.482	-.367

Table 1. Individual Pearson correlation coefficients from each run of the four simulations.

Table 1 lists the Pearson correlation coefficients for each run from the four simulations of 10 runs each. From these we calculated the mean correlation coefficient and standard deviation (in brackets) between tolerance and selectivity: (1) -.41 [.119]; (2) -.39 [.123]; (3) -.45 [.105]; and (4) -.42 [.160]. These results closely fit the correlations in Zoccolan et al. (2007); namely, -.39.

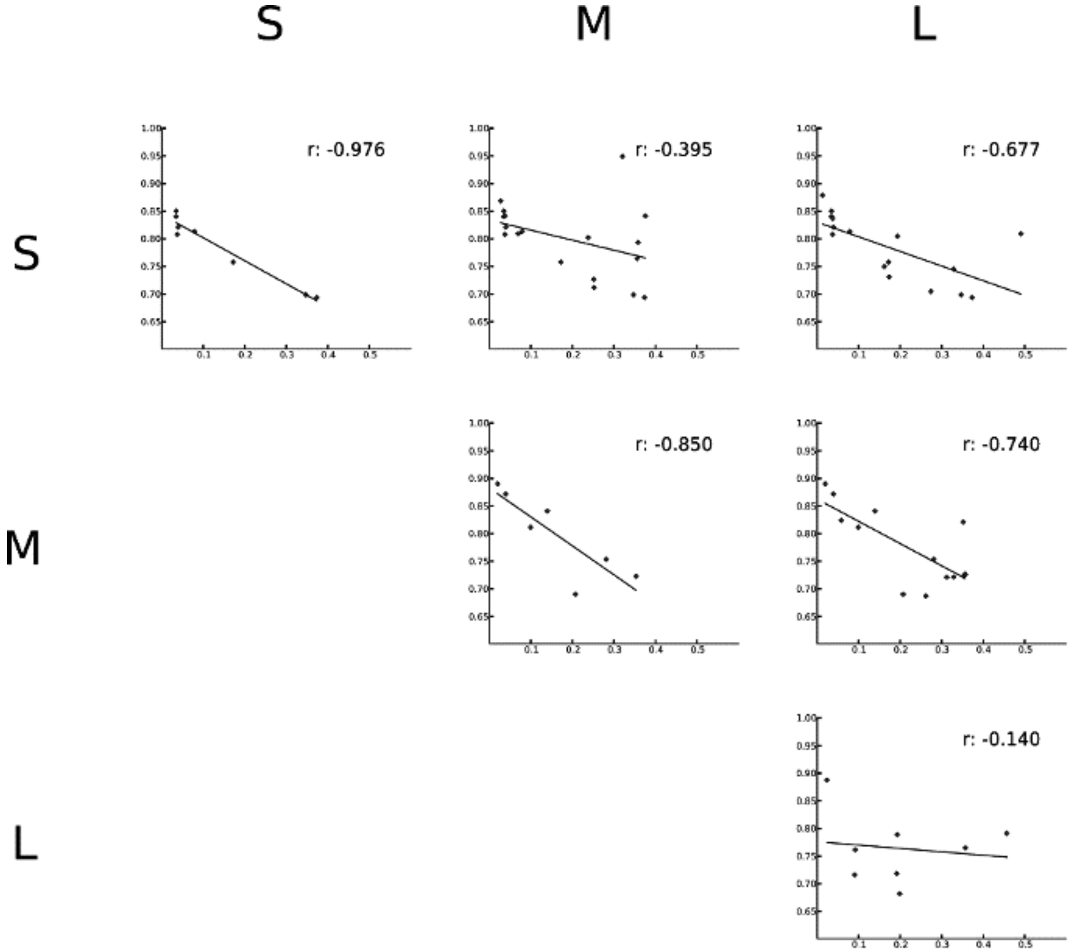

Figure 4. Simulations of Zoccolan et al. (2007) with various scales removed. The graphs are arranged in a grid, where S = small, M = medium, and L = large. Simulations at matching pathways utilize a single scale; e.g., the small-scale simulation is located at row S column S, while the small and medium scales simulation can be found at row S column M. The inclusion of broader scales and the exclusion of finer ones leads to the reduction in the tradeoff.

Simulations with subsets of scales. Key model mechanisms were perturbed to probe their effect on the tradeoff. First, simulations were conducted with only a subset of the full range of scales. The results are summarized in Figure 4. They quantify the qualitatively plausible hypothesis that the smaller scale cells are more selective and less tolerant, and vice versa for the larger ones.

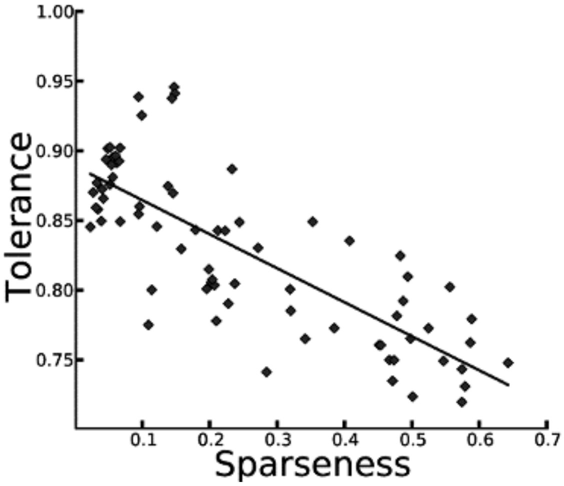

Figure 5. Simulation of Zoccolan et al. (2007) without the log-polar transform. The general tradeoff is not affected, though the slope of the regression changes dramatically.

Simulations without cortical magnification. Simulations were also carried out without the cortical magnification factor; see Figure 5. This resulted in mean correlation coefficients of -.645 and -.676 with standard deviations of .083 and .06 for the first and second set of simulations, respectively (Table 2). Since these simulations retained multiple scales and the type of classifier used to simulate IT, it is plausible that they could generate the tradeoff. Hence, the absence of cortical magnification, similar to the loss of particular scales, significantly changed the degree of correlation without changing the general tradeoff. Further simulations were conducted to better understand the factors that generated the quantitative features of the tradeoff that were described in the Zoccolan et al. (2007) data.

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
Sim. 1	-.756	-.571	-.738	-.585	-.466	-.643	-.666	-.689	-.632	-.705
Sim. 2	-.569	-.694	-.692	-.694	-.667	-.777	-.674	-.651	-.589	-.748

Table 2. Results of simulations conducted without the log-polar transform.

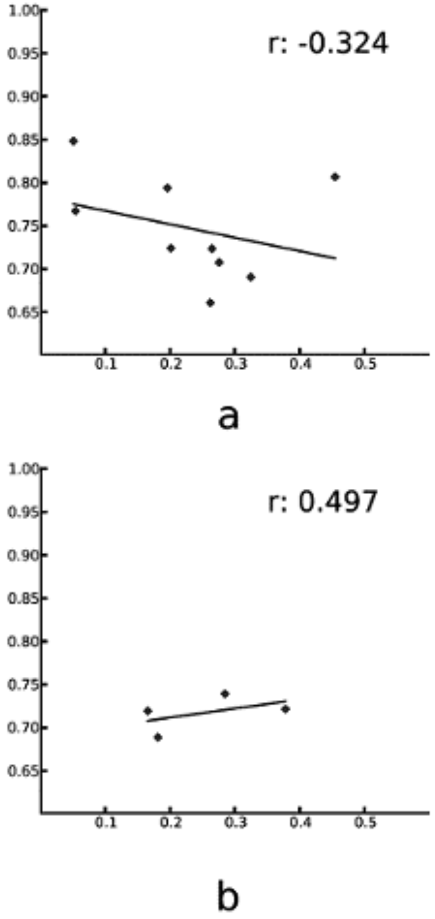

Figure 6. (a) Simulation of Zoccolan et al. (2007) using only the large scale with foveal and parafoveal input. (b) The same simulation using only parafoveal input.

Simulation with only parafoveal inputs. Simulations were run where the network was trained only on images presented parafoveally—both above and below the centered fovea. Figure 6b shows the dramatic effect of training on parafoveal images for a network that used only the largest scale. In this case, the tradeoff was either substantially changed or eliminated altogether. That is, neither very selective nor very tolerant cells formed. Taking a closer look at the weight matrices (that is, the learned adaptive weights in pathways between the preprocessed input patterns and the learned recognition categories) revealed that the average number of non-zero weights was higher when training with both foveal and parafoveal data (Figure 7).

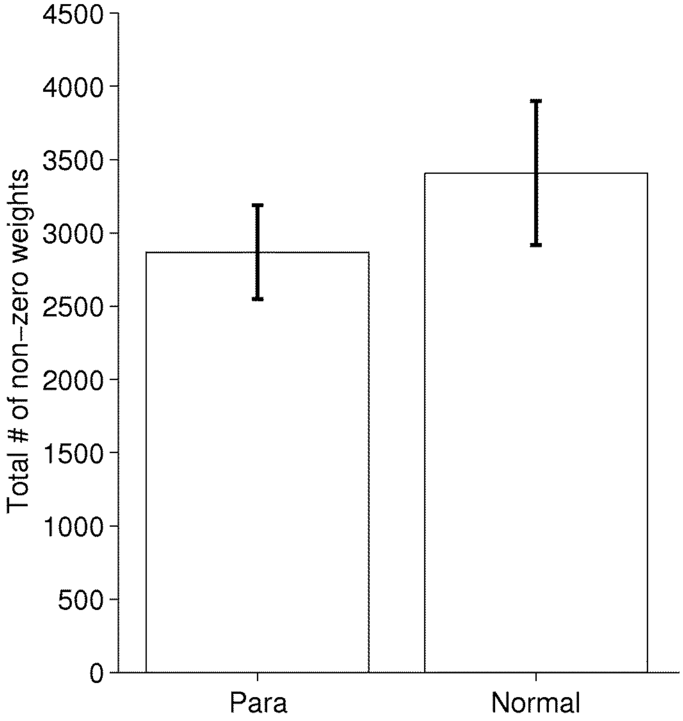

Figure 7. Comparison of the number of non-zero weights in the simulations shown in Figure 6. The *Para* case shows the average number of non-zero weights per category cell when the network was trained with parafoveal input only. *Normal* is the case in which the network was trained using both foveal and parafoveal input.

The effect of the number of non-zero weights was teased apart by then looking again at the weight matrices of the simulations shown in Figure 4. As shown in Figure 8, strong negative and positive correlations occur, respectively, between selectivity (Figure 8a) and tolerance (Figure 8b) and the number of non-zero weights.

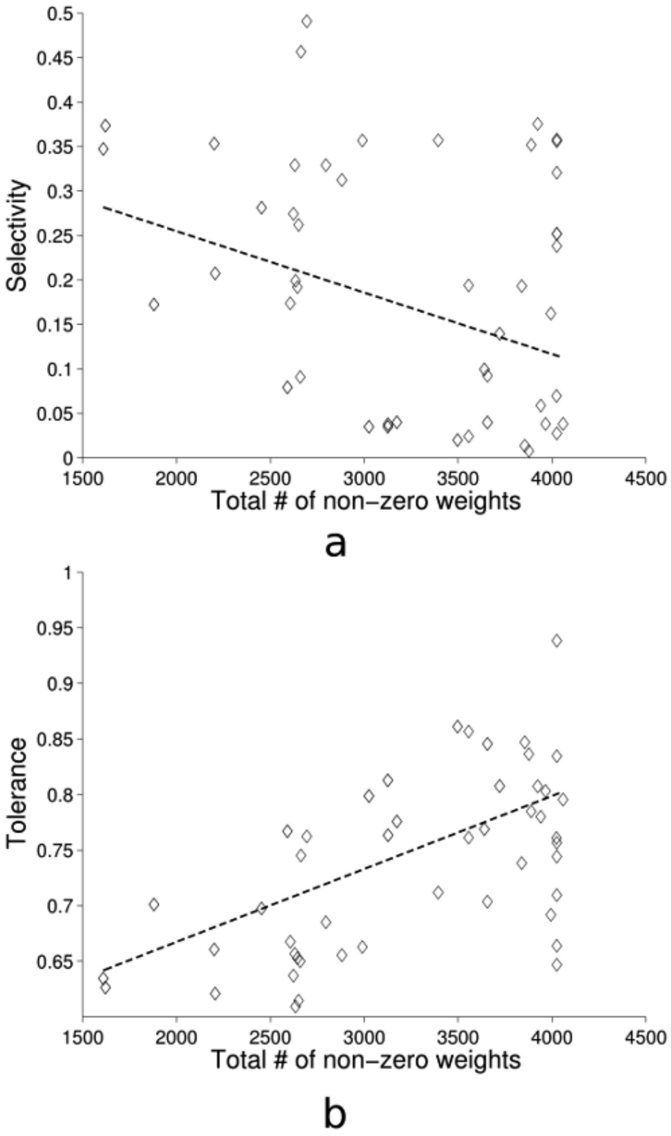

Figure 8. Analysis of data pooled from simulations shown in Figure 4. (a) Category node selectivity as a function of the total number of non-zero weights (r = -.38, p < .002). (b) Category node tolerance as a function of the total number of non-zero weights (r = .64, p < .0001).

Parafoveal simulation without cortical magnification. A simulation was also run without cortical magnification using only parafoveally presented images and exhibited a strong tradeoff between selectivity and tolerance, even when using only single scales, with an average correlation coefficient of −.795 (standard deviation .099). This result contrasts markedly with the lack of a tradeoff when parafoveal images are processed by a single scale when the cortical magnification factor was intact (Figure 6b).

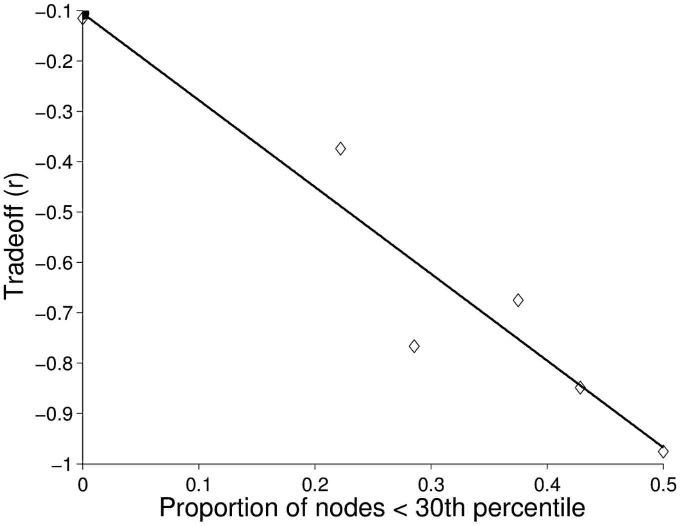

Figure 9. A scatter plot of the tradeoff as a function of the proportion of the number of non-zero weights less than the 30^{th} percentile, using data from the simulations shown in Figure 4 (r = -.9012, p < .02). There were zero nodes whose number of non-zero weights was less than zero in the simulation using only parafoveal input. The linear regression shown here predicts that there would be a smaller tradeoff, and this was indeed the case.

What model mechanisms can explain this pattern of simulation results? We found that the standard deviation in the number of non-zero learned weights abutting category cells was much higher in the parafoveal simulation without the cortical magnification factor (491.99) than in the parafoveal simulation that included it (321.23). Additionally, the standard deviation in the number of non-zero weights for all of the simulations shown in Figure 4 is relatively high (761.74). This finding seems to point to a correlation between the standard deviation of the learned weights and the tradeoff. On the other hand, such a correlation does not fully explain the variance in standard deviation among simulations in Figure 4 (data not shown). A more detailed analysis of the weights was therefore performed.

First, a significant negative correlation was found between the number of non-zero weights and the selectivity of individual category nodes (r = -.3822, p < .001) and a positive correlation between the number of non-zero weights and tolerance (r =.6427, p = 0) in the simulations shown in Figure 4. Then, a highly significant and strong negative correlation was found between the proportion of nodes whose total number of non-zero weights was below the 30^{th} percentile (for all simulations shown in Figure 4) and the correlation value of the tradeoff (r = -.9012, p < .02), as shown in Figure 9. The presence of more selective cells with fewer non-zero weights increased the magnitude of the tradeoff (i.e., led to the lower r value in the correlation between selectivity and tolerance). By the same measure, this would predict a minimal tradeoff in the simulation with parafoveal data presented with the log-polar transform, in which there were no nodes in which the number of non-zero weights was below the 30th percentile. Indeed, this was the case.

The specific model mechanism responsible for this is the process of category selection and learning by the model's ART classifier. In essence, if a given input feature pattern does not match any top-down expectation projecting to a model IT cell, then a new cell forms with weights that match the input pattern. In the presence of high-resolution natural images, as in most of the simulations with the log-polar transform using foveal input, and in the simulations without the log-polar transform, the varying statistics of natural images led to the creation of category cells with a high dispersion in the number of non-zero weights. In this case, the network will learn highly tolerant and selective cells. Additionally, the specific form of the ART bottom-up adaptive filter (see Appendix Equation (A15)) results in the contrasting correlations between the number of non-zero weights and selectivity and tolerance. More specifically, a category node with a higher number of non-zero weights is likely to be partially activated by many stimuli, since it is likelier to find a partial match for a larger template, leading to high tolerance. On the other hand, by decreasing the number of non-zero weights, one decreases the chances that a particular category node will be partially activated by a large selection of stimuli.

To summarize, the model generates the tradeoff between selectivity and tolerance without either multiple spatial scales or the log-polar transform, yet does not generate this tradeoff when presenting images parafoveally with the log-polar transform, akin to the reduction in the tradeoff when presenting images foveally and parafoveally and using only the largest spatial scale (Figure 4 bottom right). These two cases are similar in that coarse information was presented to the model classifier, and the distribution of learned weights is homogeneous. Thus, the tradeoff results from the variety in the number of non-zero weights across model IT cells, which is caused by: (1) the presentation of natural images at high-resolution (at the fovea in the case of the log-polar simulations), and (2) the classifier used to form model IT cells, which learns weights that are highly dispersed given complex, high-resolution input. Moreover, the interaction of multiple spatial scales and the log-polar transform, instead of causing the tradeoff, merely changes its degree.

Akrami et al. (2009) data simulations. The model simulations of neurophysiologically recorded ITa cells to image morphs were carried out as follows. Since Akrami et al. (2009) employed proprietary software to compute the morphs, our procedure used a progressive alpha blend from one image to the other; that is, the first image in a pair was made progressively more transparent while the second image was made less transparent until only the second image was visible. As in the Zoccolan et al. (2007) simulations, natural images were taken from the Cal Tech 101 database. To train the model, 16 images were used for learning (see Figure S2 in the Supplemental Information). Then learning was frozen and the model was tested on eight morph pairs made from the training stimuli, similar to the procedure used by Akrami et al. (2009), leading to the model fit of the data in Figure 1d. Moreover, it should be noted that we focused our model simulations on the first 100-200 ms of an IT cell's response after the presentation of an image.

The graded response to the image morphs, from the 'ineffective' to the 'effective' image, results from how the information in multiple spatial scales is classified. A model IT cell that codes for a chair will respond to another object commensurate to the degree it resembles the same chair. This yields a graded response when presenting a series of morphs from, for instance, a tennis ball to the chair. Moreover, some model IT cells form their category representation based on a coarse scale and others at a fine scale, resulting in a population of both selective and non-selective cells. Since the input to the model IT is also presented at multiple spatial scales, certain cells will have a highly graded response and others a comparatively flat response to a series of morphs. The average of these two groups provides a close match to the data from Akrami et al. (2009), which is itself a population average of extracellularly recorded ITa cells.

Similar to the Zoccolan et al. (2007) simulations, we lesioned particular scales in the model visual pathway to test their individual contributions. As one might expect, the graded response to image morphs was invariant to the scale(s) used, which is summarized in Figure 10.

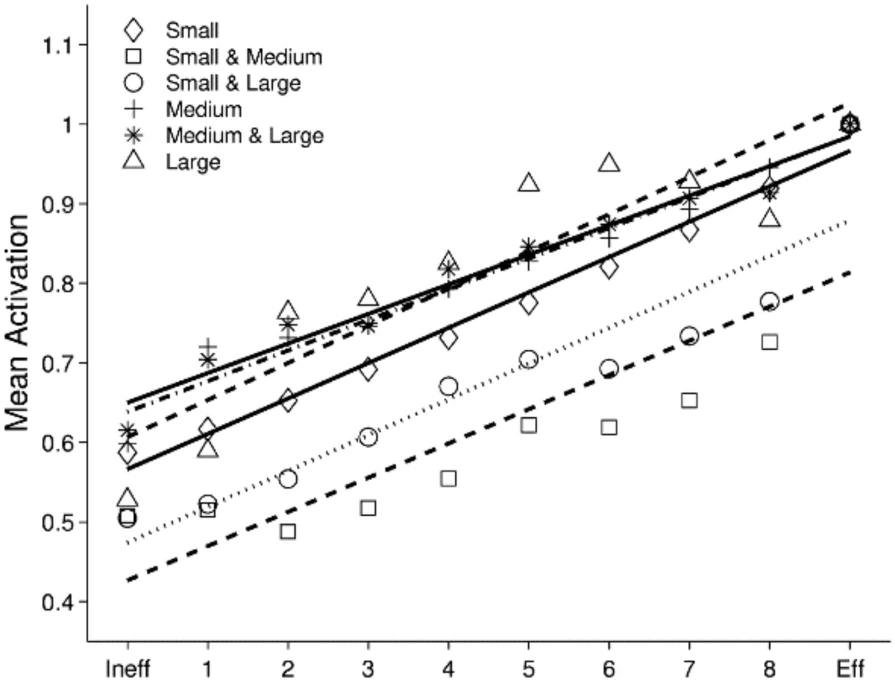

Figure 10. Simulations of Akrami et al. (2009) under lesions of particular scales in the model visual pathway. Notice that the general trend is invariant to the scale(s) used.

Simulations with different vigilance parameters. Figure 11a shows how the model responds to the Zoccolan et al. (2007) experimental conditions under high vigilance ($\rho = .9$), whereas Figure 11b shows how the model responds to the Akrami et all. (2009) experimental conditions under lower vigilance ($\rho = .2$). These simulation results differ from those in Figure 1b and 1d, respectively, because of the way the bottom-up adaptive weights, or long term memory traces, that project to model IT cells learn category representations. In the case of the Zoccolan et al. (2007) data, the tradeoff effectively disappears due to the development of model IT cells that are both selective and tolerant. The use of high vigilance drives the formation of highly selective cells, while categories formed at coarse spatial scales are more tolerant than those formed at fine scales. On the other hand, the use of low vigilance for the Akrami et al. simulation led the model IT cells to learn coarser categories, even at fine spatial scales. Hence, a number of cells responded similarly to each object, leading to a flatter response to a series of image morphs.

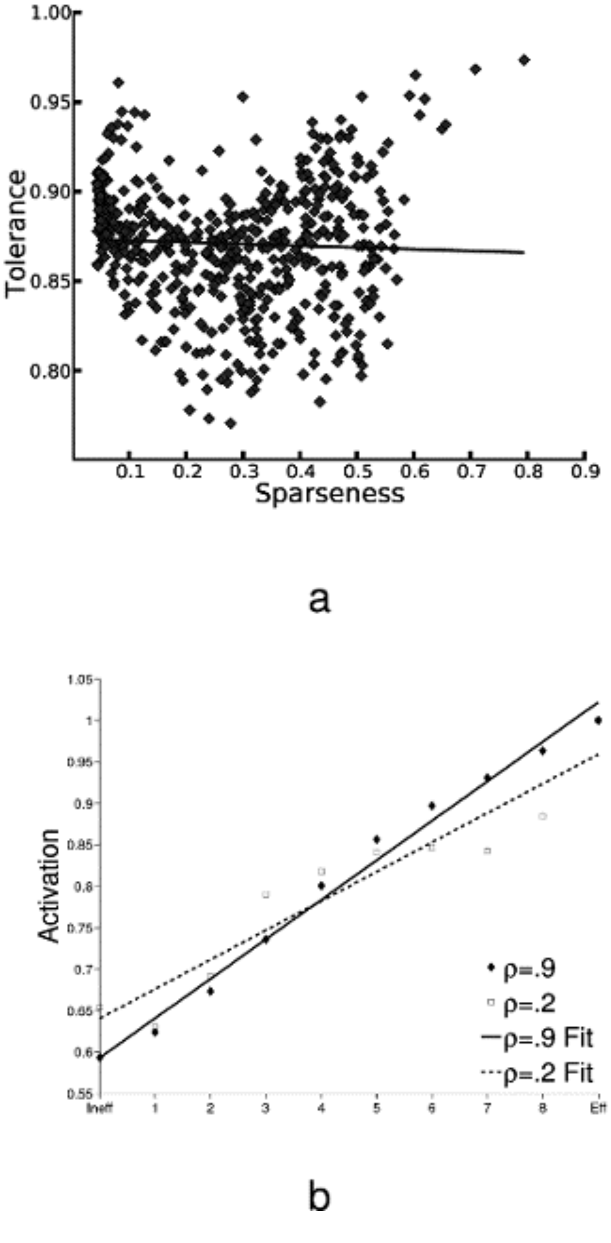

Figure 11. (a) Model simulation of Zoccolan et al. (2007) experiment with high vigilance ($\rho = .9$). (b) Results from model simulation of Akrami et al. with low ($\rho = .2$) and high ($\rho = .9$) vigilance. When vigilance is low, the model produces a flatter response to the image morphs (see main text for details). The residual sum of squares to the linear fit is .0024 for high vigilance and .0117 for low vigilance, indicating that high vigilance leads to a more linear trend.

Discussion: Reconciling Invariant Recognition with the Tradeoff

Emergent properties of a cortical hierarchy. Our model shows how, when particular combinations of known brain processes interact together in a specific order, they lead to emergent properties of model cortical cells that can explain data which are otherwise difficult to understand. Various alternative models incorporate some of these processes. For example, multiple spatial scales have been used for both technical applications in object recognition (Burt & Adelson, 1983; Porat & Zeevi, 1988; Rosenfeld & Thurston, 1971) and to simulate aspects of object recognition by the What cortical processing stream (Bradski & Grossberg, 1995; Fukushima, 1980, 1988; Riesenhuber & Poggio, 1999; Serre, Oliva, & Poggio, 2007; Watson, 1987). Our model uses multiple-scale filtering with Gaussian kernels (most others employ Gabor filters) to account for the heterogeneity of receptive field sizes throughout the visual neocortex, from V1 to ITa. Our model also adopts the common approach of using a classifier after this multiple-scale filtering to simulate the computations of ITa and the PFC (Fukushima, 1980, 1988; Gochin, 1994; Grossberg & Williamson, 1999; Riesenhuber & Poggio, 1999; Serre et al., 2007). Although qualitative aspects of the Zoccolan et al. (2007) tradeoff can be simulated without a number of model components (see Figures 4 and 5), taken all together they allowed us to fit the data quantitatively. In particular, we are unaware of other category learning models that include the log-polar transform, or cortical magnification factor, a key transformation of information between the LGN and V1, with the exception of Bradski & Grossberg (1995) and more recently Fazl et al. (2009), although some others have used a simple approximation (Watson, 1987) to simulate cortical magnification.

Cortical magnification and multiple scales. The cortical magnification factor and the use of multiple scales provide important means by which to control the degree of the tradeoff; that is, they alter the magnitude of the correlation between selectivity and tolerance, but not its sign. In the model, the dispersion of non-zero weights underlies the tradeoff, and this is a factor controlled by the Adaptive Resonance Theory, or ART, category learning circuit that we to model IT. In the presence of high-resolution natural images, the variance in the complex, natural input stimuli drives the formation of model IT cells with high variability in the number of non-zero weights. Then the bottom-up ART adaptive filter leads to a negative correlation between number of non-zero weights and selectivity, and a positive correlation between the number of non-zero weights and tolerance (Figure 8), as demonstrated through the analysis carried out on the simulations shown in Figure 4.

Vigilance-modulated attentive matching. A critical part of the model is thus the use of an ART classifier wherein attentive matching between bottom-up feature patterns and learned top-down expectations solves the *stability-plasticity dilemma*, and in so doing clarifies the role of task-selective vigilance in enabling the same category learning circuit to learn concrete and specific recognition categories, as well as abstract and general recognition categories. Such an attentively-modulated classifier circuit is qualitatively different from the feedforward classifiers that are used in many current models (e.g., Riesenhuber & Poggio, 1999) and that are commonly used in the linear classifiers of many machine learning applications (Serre et al., 2007).

These attentive matching task-selective processes enable our model to quantitatively simulate neurophysiological data concerning how ITa cells may generate both the selectivity/tolerance tradeoff (Zoccolan et al., 2007) and graded responses to image morphs (Akrami et al., 2009). Significantly, the selectivity/tolerance tradeoff was recorded during passive viewing conditions and thus seems to represent a low vigilance task, whereas the image morph experiment required active discrimination during a delayed match-to-sample task and is thus a high vigilance task.

A possible link to mental disorders. As noted in the Model Overview, Grossberg & Versace (2008) have proposed how vigilance may be controlled by novelty-sensitive corticothalamic mismatches that activate the nonspecific thalamus, which activates acetylcholine release via the nucleus basalis of Meynert, thereby increasing excitability of cortical layer 5 cells, and leading to cortical reset in layer 4. Deficient vigilance control has been predicted to occur in autism and medial temporal amnesia, among other mental disorders (Carpenter & Grossberg, 1993; Grossberg & Seidman, 2006). If these anatomical and pharmacological predictions about vigilance control are confirmed, then in addition to setting the stage for providing a direct test of whether vigilance is high in the Zoccolan et al. (2007) paradigm and low in the Akrami et al. (2009) paradigm, then manipulations that cause vigilance to be fixed at high or low levels may help to create an animal model of these disorders.

Linking What and Where during view-invariant category learning. The passive viewing conditions during the Zoccolan et al. (2007) experiment lead to several questions about what might happen in variants of this task that are done under high vigilance. In this regard, during active scanning of objects with saccadic eye movements, spatial attentional mechanisms in the parietal cortex may modulate object category learning in ITa to enable binding of multiple object views into a more position- and view-invariant object category representation by incremental learning during free viewing of a scene with eye movements. Fazl, Grossberg & Mingolla (2009) have developed the ARTSCAN model to explain how view-invariant object categories may be learned using such spatial attentional modulation. Cao, Grossberg, & Markowitz (2010) have extended this analysis to predict how view-, position-, and size-invariant categories may be learned. Fazl et al. (2009) began this analysis by simulating how spatial attention in parietal cortex may fit itself to an object's visual form. Such an "attentional shroud" enables an emerging view-invariant object category to remain active as view categories are associated with it one at a time. Multiple view categories may be reset to enable the next view category to be activated and associated with the view-invariant category. A shroud enables the view-invariant category to remain active, despite the reset of its view categories, by inhibiting the reset mechanism that would otherwise inhibit the view-invariant object category.

When a shift in spatial attention occurs, an active shroud collapses and elicits a transient reset burst, also in the parietal cortex, that inhibits the view-invariant object category in the temporal cortex, and thereby enables a shift in categorization rules to occur when the next object is attended. The transient reset signal and the attentional shroud are model instantiations of processes that are often heuristically called transient and sustained attention. This model prediction has received support from data of Chiu & Yantis (2009), who have reported the existence of a transient signal within the parietal cortex that mediates correlated spatial attention shifts and shifts in categorization rules. Thus the ARTSCAN model predicts how spatial attention in the Where cortical processing stream may regulate invariant object category learning in the What cortical processing stream.

The ARTSCAN model, just like the current model, incorporates log-polar boundary processing followed by attentive vigilance-modulated ART category learning. It does not, however, make use of multiple spatial scales. Instead, ARTSCAN emphasizes how spatial attentional processing can modulate how view-invariant categories are learned. This observation, when taken with the fact that the Zoccolan et al. (2007) experiment was done using passive viewing, suggests the utility of carrying out additional neurophysiological experiments that test the selectivity/tolerance tradeoff both before and after active scanning and view-invariant object learning take place. In particular, if the Zoccolan et al. (2007) task is altered to require more focal attention to the stimuli, will results similar to those in Figure 11a be found? More generally, simultaneous electrode studies are much to be desired in which correlated activations of spatial attention in parietal cortex and of category-representing cells in inferotemporal cortex directly test how spatial attention in the Where cortical stream influences the course of category learning in the What cortical stream. In particular, how do the transient bursts reported by Chiu & Yantis (2009) mediate between spatial attention and category learning? Such studies can clarify how the brain bridges between the What and Where cortical streams to coordinate spatial and object attention and eye movement search to enable invariant category learning and recognition of an ever-changing world.

APPENDIX: Mathematical Model

Input Images

The complete dataset of images for the Zoccolan et al. (2007) and Akrami et al. (2009) simulations are shown in Figures S1 and S2 in the Supplemental Information, respectively.

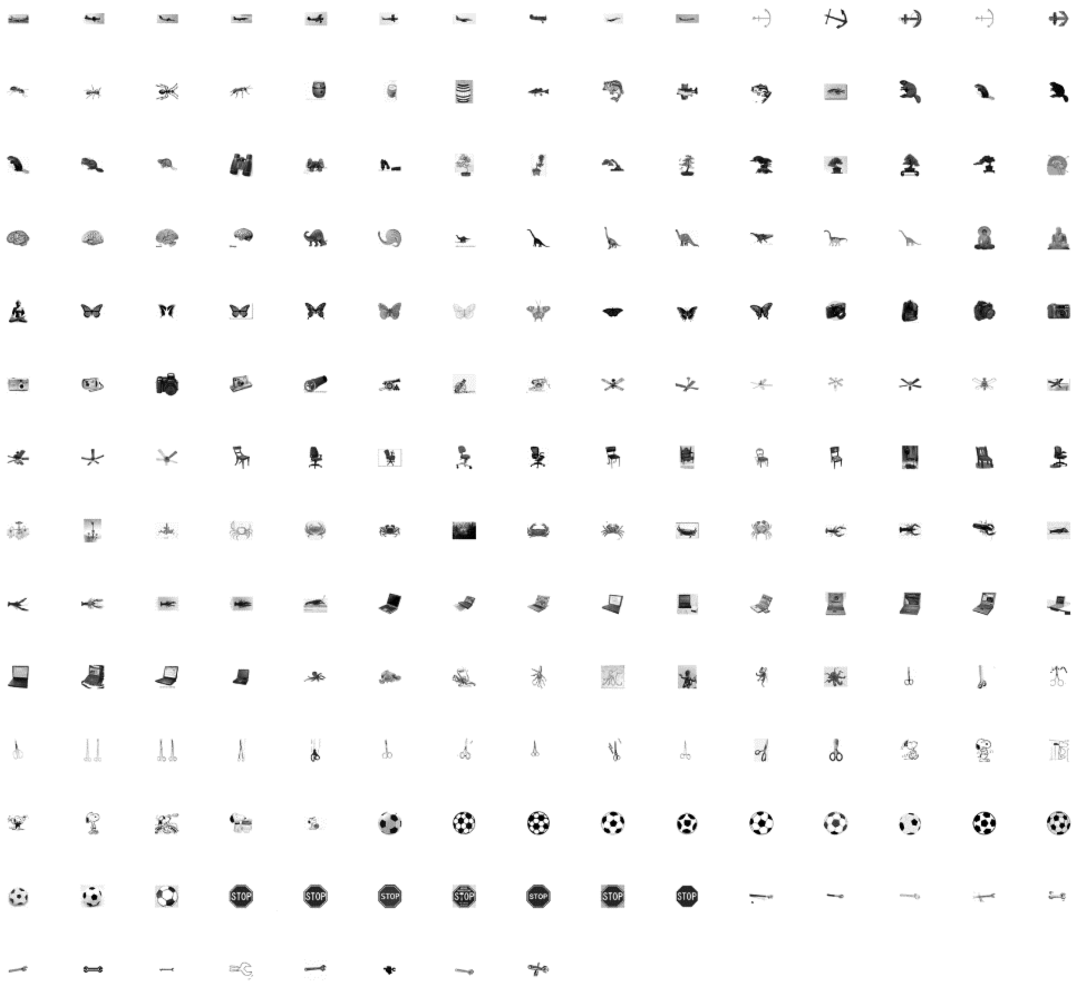

Figure S1. All 203 images were adapted from the Cal Tech 101 database, as were the stimuli used in the Zoccolan et al. (2007) experiment.

Stimulus Set and Simulation Procedure

To simulate the Zoccolan et al. (2007) data, a master dataset of images was constructed from the Cal Tech 101 database (Figure S1) scaled to approximately 100 x 100 pixels and presented against a uniform white background of 300 x 300 pixels (Fei et al., 2006). Images were superimposed over the background at: (1) the center, (2) 50 pixels below the center, or (3) 50 pixels above the center. This resulted in a master dataset of 609 images (203 images x 3 positions). For the *learning phase* of each run, 200 images were randomly drawn from the master dataset without replacement and new model IT recognition categories were incrementally learned by a variant of Adaptive Resonance Theory (ART), called fuzzy ART, that is described below. During the *testing phase*, learning was frozen while presenting two fixed subsets of the master dataset: (1) all 203 objects presented at the center to measure selectivity and (2) 49 objects presented at all three positions to measure position tolerance (49 objects x 3 positions = 147 total images). In both phases, each input image was presented to the network one at a time. In the Akrami et al. (2009) simulations, natural images were again taken from the Cal Tech 101 database. To train the model, we used 16 images for learning (see Figure S2). Then, we froze learning and tested the model on 8 morph pairs made from the training stimuli, similar to the procedure used by Akrami et al. (2009).

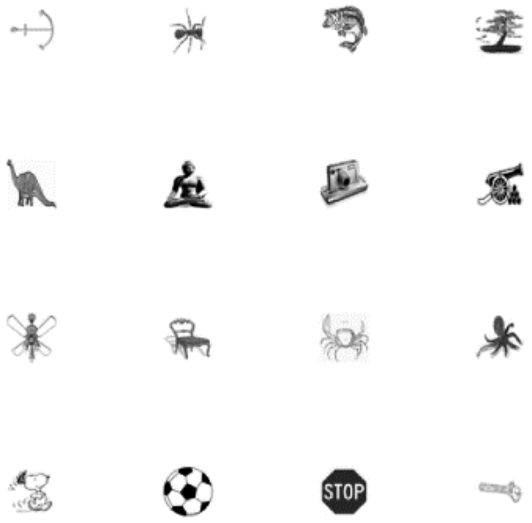

Figure S2 . Stimuli used to train the model for the Akrami et al. (2009) simulation.

Model Equations

Preprocessing. Each input image is a circular 'cut-out' of the original image, which is then split into oversampled hemi-retinas (Figure 3). Oversampling includes part of the opposing hemi-retina near the vertical meridian, consistent with neurophysiological evidence (Van Essen et al., 1984). This process is described in further detail in Fazl et al. (2009). Cell processing begins by taking the average of the three RGB color values of an input image, split into hemi-retinas:

$$I_{pq}^{h} = \frac{1}{3}(I_{pq}^{hR} + I_{pq}^{hG} + I_{pq}^{hB}) \tag{A1}$$

Here, the indices *p* and *q* correspond to the Cartesian x and y coordinates, respectively, of the input image and are used as dummy indices below; *h* indexes the hemi-retina {left, right}; and *R,G,B* the color channel.

Retina/LGN Processing. The retinal and LGN cells carry out *contrast normalization* of the input pattern using, at each of three scales (g = 1,2,3), an on-center off-surround network whose cells obey membrane, or shunting, equations (Grossberg, 1973). The narrow on-center is defined by a single pixel and the off-surround by a Gaussian kernel, centered on location *(i,j)* in the image, whose breadth varies with the scale (Bhatt et al., 2007; Grossberg & Hong, 2006). Solved at equilibrium, the activity, or potential, x_{ij}^{hg+}, of the ON cell at position (*i,j*), hemi-retina h, and scale *g* is:

$$x_{ij}^{hg+} = \frac{I_{ij}^{h} - \sum_{p,q} S_{ijpq}^{g} I_{pq}^{h}}{1 + I^{h}_{ij} + \sum_{p,q} S_{ijpq}^{g} I_{pq}^{h}}, \tag{A2}$$

where the scale-dependent Gaussian off-surround kernels are defined by:

$$S_{ijpq}^{g} = \frac{1}{2\pi\sigma_{sg}^{2}} \exp\left[-\frac{(i-p)^2 + (j-q)^2}{2\sigma_{sg}^{2}} \right]. \tag{A3}$$

In (A2), the breadth of the three scales is determined by $\sigma_{sg} = (1,2,3)$ in (A3). The denominator in (A2) reflects shunting dynamics and carries out contrast normalization. The outputs of the ON and OFF channels are:

$$\begin{aligned} X_{ij}^{hg+} &= [x_{ij}^{hg+}]^{+} \\ X_{ij}^{hg-} &= [-x_{ij}^{hg+}]^{+}, \end{aligned} \tag{A4}$$

where $[x]^{+} = \max(x,0)$ denotes a half-wave rectifying output signal, and the superscripts + and – refer to the ON and OFF channels, respectively.

Log-polar Map. The ON and OFF channels of each hemi-retina undergo a log-polar transform that maps from retina to cortex, with the left (right) hemi-retina forming the right (left) hemispheric image:

$$\begin{aligned} z^{hc} &= re^{i\theta} \\ w^{hc} &= \log(z^{hc} + a). \end{aligned} \tag{A5}$$

In (A5), z is a complex number formed by the retinal image in polar coordinates, *h* refers to the hemi-retina, *c* labels the ON and OFF channels, and *a* is a constant (.7). The output of this operation is a log-polar map of each hemi-retina, $V^{left,c}$ and $V^{right,c}$. Both of the hemi-retinas are concatenated into a single image W^{c}, where *c* designates either the ON channel (c+) or the OFF channel (c-):

$$W^{c+} = (V^{left,c+}, V^{right,c+}) \tag{A6}$$

$$W^{c-} = (V^{left,c-}, V^{right,c-}). \tag{A7}$$

Simple and Complex Cell Boundary Processing. Model V1 simple cells are activated by LGN ON and OFF outputs that are filtered by spatially elongated and offset Gaussian kernels. In particular, simple cell activity y^g_{ijk} at position (*i,j*) , orientation *k*, and scale *g* obeys the shunting on-center off-surround equation (Bhatt et al., 2007):

$$\begin{aligned}\frac{d}{dt}y^g_{ijk} = -\alpha y^g_{ijk} + (1-y^g_{ijk})\sum_{p,q}\left(W^{c+}_{pq}G^{g+}_{pqijk} + W^{c-}_{pq}G^{g-}_{pqijk}\right) \\ -(1+y^g_{ijk})\sum_{p,q}\left(W^{c+}_{pq}G^{g-}_{pqijk} + W^{c-}_{pq}G^{g+}_{pqijk}\right).\end{aligned} \tag{A8}$$

In (A8), the passive decay rate $\alpha = 1$. In the excitatory term of (A8), the log-polar transformed LGN ON cell output signals W^{c+}_{pq} are filtered by the oriented, spatially-elongated Gaussian kernel G^{g+}_{pqijk}, while the LGN OFF output signals W^{c-}_{pq} are filtered by a similar kernel G^{g-}_{pqijk}. The centers of the kernels G^{g+}_{pqijk} and G^{g-}_{pqijk} are offset in mutually opposite directions from each simple cell's centroid along an axis perpendicular to the simple cell's direction of elongated sampling. In the inhibitory term of (A8), the same kernels sample an LGN channel complementary to the one in the excitatory term. The net activity of model simple cells is thus a measure of image feature contrast in its preferred orientation. In mathematical terms, the kernels in (A8) are:

$$\begin{aligned}G^{g+}_{pqijk} = \frac{1}{2\pi\sigma_{lg}\sigma_{sg}}\exp\Bigg(-\frac{1}{2}\Bigg\{\left[\frac{(p-i+m_k)\cos\left({}^{\pi k}\!/_{4}\right)-(q-k+n_k)sin\left({}^{\pi k}\!/_{4}\right)}{\sigma_{lg}}\right]^2 \\ +\left[\frac{(p-i+m_k)\sin\left({}^{\pi k}\!/_{4}\right)+(q-j+n_k)\cos\left({}^{\pi k}\!/_{4}\right)}{\sigma_{sg}}\right]^2\Bigg\}\Bigg) \\ G^{g-}_{pqijk} = \frac{1}{2\pi\sigma_{lg}\sigma_{sg}}\exp\Bigg(-\frac{1}{2}\Bigg\{\left[\frac{(p-i-m_k)\cos\left({}^{\pi k}\!/_{4}\right)-(q-k-n_k)sin\left({}^{\pi k}\!/_{4}\right)}{\sigma_{lg}}\right]^2 \\ +\left[\frac{(p-i-m_k)\sin\left({}^{\pi k}\!/_{4}\right)+(q-j-n_k)\cos\left({}^{\pi k}\!/_{4}\right)}{\sigma_{sg}}\right]^2\Bigg\}\Bigg).\end{aligned} \tag{A9}$$

Here, g denotes scales 1, 2, and 3, and the plus and minus signs indicate spatial shifts in opposite directions and the index *k* indicates a given orientation. The long-axis variance $(\sigma_{l1},\sigma_{l2},\sigma_{l3}) = (3/4, 9/4, 27/4)$, the short-axis variance $(\sigma_{s1},\sigma_{s2},\sigma_{s3}) = (1/4, 3/4, 9/4)$, and the offset vector $(m_k, n_k) = \left(\sin\left(\pi k/4\right), \cos\left(\pi k/4\right)\right)$. Finally, the output signals were half-wave rectified:

$$\begin{aligned} Y_{ijk}^{g+} &= \left[y_{ijk}^{g}\right]^{+} \\ Y_{ijk}^{g-} &= \left[-y_{ijk}^{g}\right]^{+}. \end{aligned} \tag{A10}$$

Each V1 complex cell receives activity from pairs of simple cells at the same position that are selective to opposite polarities and the same orientation. It hereby acts as an oriented filter that pools over opposite contrast polarities. The complex cell activity z_{ijk}^{g} for position (*i,j*), orientation *k*, and scale *g* obeys the equation:

$$z_{ijk}^{g} = Y_{ijk}^{g+} + Y_{ijk}^{g-}, \tag{A11}$$

where Y_{ijk} is the activity described in (A10), and the superscripts g+ and g- indicate opposite contrast polarities. This computation yields twelve ‘hemispheric’ images, which are the outputs of boundary processing mechanisms via simple and complex cells in three spatial channels and four different orientations at each position.

To compute an unoriented boundary strength at each position and scale in model V2, the boundary images are summed across orientation:

$$Z_{ij}^{g} = \sum_{k} z_{ijk}^{g}. \tag{A12}$$

To prepare the data for IT learning and recognition, $Z^{g}{}_{ij}$ is flattened from a two-dimensional matrix to a row vector. That is, the matrix of M rows and N columns is mapped to an M x N element row vector $\vec{J}^{g}$. Thus, the value in a given row *m* and column *n* is mapped to element *N*(*m-1)+n* of $\vec{J}^{g}$; e.g., the value at row 2 and column 2 in a 12 column matrix maps to element 14 of $\vec{J}^{g}$. The vector $\vec{J}^{g}$ is then normalized:

$$I_{i}^{g} = \frac{J_{i}^{g}}{max_{k}(J_{k}^{g})} \tag{A13}$$

via an operation that emulates a shunting network with global inhibition. This fixes all values of the normalized vector $\vec{I}^{g}$ within the interval [0,1]. Equations (A1)-(A13) are computed for all images, thus preparing three ‘flattened’ vectors (for three spatial scales) for each input image. These three normalized input vectors of different scale are then categorized by model IT.

ITa Recognition. Category learning by IT was modeled using a variant of Adaptive Resonance Theory that is called fuzzy ART (Carpenter, Grossberg, & Rosen, 1991), which is realized as an algorithm for ease of implementation. How variants of these algorithmic operations may be embedded in laminar thalamocortical circuits is modeled in Grossberg and Versace (2008). Each input is an M x N-dimensional 'flattened' boundary vector $\vec{I}$. Each category (*j*) is activated via a bottom-up adaptive filter that learns a vector $\overrightarrow{w_j} = (w_{j1},...,w_{jm})$ of adaptive weights or long-term memory (LTM) traces. The number of potential categories $N(j = 1,...,N)$ is arbitrary. At first,

$$w_{j1} = ...w_{jM} = 1 \tag{A14}$$

and each category is then said to be *uncommitted.* After a category is selected for coding, it becomes *committed.* Each LTM trace w_{ji} is monotone non-increasing through time and so converges to a limit. This property assures the stability of learned memories. In Fuzzy ART, both the bottom-up and top-down learned weights are the same, so a single weight vector $\overrightarrow{w_j}$ corresponding to each learned category suffices to represent both bottom-up and top-down learning. Dynamics are determined by a choice parameter $a > 0$, a learning rate parameter $\beta \in [0,1]$, and a vigilance parameter $\rho \in [0,1]$. Each boundary input vector $\vec{I}$ is processed by a bottom-up adaptive filter that computes a choice function (Carpenter & Gjaja, 1994) T_j with which to select each category j:

$$T_j(\vec{I}) = \frac{|\overrightarrow{I_g} \wedge \overrightarrow{w_j}|}{\alpha + |\overrightarrow{w_j}|}, \tag{A15}$$

where the fuzzy AND operator $\wedge$ is defined by

$$(x \wedge y)_i = min(x_i, y_i) \tag{A16}$$

and where the norm $|\cdot|$ is

$$|x| = \sum_{i=1}^{M} |x_i|. \tag{A17}$$

The fuzzy AND operator may be interpreted in terms of the fraction of learned postsynaptic sites that can be activated by each input.

After all bottom-up inputs are registered at the category level, the cells compete via long-range lateral inhibition to choose that category that receives the largest input, which is then stored in short term memory. For convenience, $T_j(I)$ may be written as T_j, and the category choice is denoted by:

$$T_J = max\{T_j : j = 1...N\}. \tag{A18}$$

In the case of a tie, the cell with the smallest index is chosen. Selection of a category enables top-down read-out of its learned expectation to the distributed feature level. As noted above, in fuzzy ART, the same adaptive weights act in the top-down learned expectation as in the bottom-up adaptive filter.

As in Figure 1, each active bottom-up input tries to turn on the orienting system, and each active cell at the distributed feature level tries to turn it off. In response to bottom-up activation alone, before the category level gets activated, there are as many active features as inputs, so the total inhibition $|\vec{I}|$to the orienting system from the distributed feature level is sufficient to shut off the excitation that is due to the distributed inputs. When the top-down expectation is active, however, it selects features that are consistent with the learned prototype, so that the total inhibition from the remaining active features is reduced to$|\vec{I} \wedge \overrightarrow{w_J}|$. Hence, in Fuzzy ART, attentive selection by biased competition uses the fuzzy AND operation which, in particular, drives a cell's response to zero if its top-down learned weight is zero.

Whether this amount of inhibition is sufficient to prevent the orienting system from being activated depends upon the *vigilance parameter* ρ*,* because each input is multiplied by ρ to generate a total excitatory input $\rho|\vec{I}|$ to the orienting system (Figure 1). Thus, a larger vigilance value makes it harder for inhibition from a poor match $|\vec{I} \wedge \overrightarrow{w_J}|$ to inhibit the orienting system, and thus easier to reset the system to search for a better matching category. Higher vigilance hereby forces learning of more concrete and specific categories.

If the total excitation is less than the total inhibition at the orienting system, then the orienting system remains quiet and allows the bottom-up and top-down signals to cycle. That is why *resonance* is said to occur if the match due to read-out of the learned expectation of an active category meets the vigilance criterion:

$$\frac{|\vec{I} \wedge \overrightarrow{w_J}|}{|\vec{I}|} \geq \rho, \tag{A19}$$

which just means that total inhibition is stronger than total excitation in the orienting system for that choice of ρ. Learning can then occur as described in equation (A21) below. If, however, total excitation exceeds inhibition at the orienting system, as described above, then the orienting system can become active and generate a novelty-sensitive arousal burst that resets the currently active category. *Mismatch reset* is thus said to occur when:

$$\frac{|\vec{I} \wedge \overrightarrow{w_J}|}{|\vec{I}|} < \rho \tag{A20}$$

Then the value of the choice function T_J in (A15) is reset to -1 for the remainder of the input presentation to prevent its persistent selection during search. In more dynamical descriptions of ART, such persistent inhibition is accomplished by an interaction of habituative transmitters with the arousal burst, which together cause selective rebounds in cell activation (i.e., the -1) that are maintained by recurrent lateral inhibition among the category cells (Carpenter & Grossberg, 1990).

The search process continues until (A19) is satisfied. Then the resonant state sets in, and the adaptive weights are updated via learning as follows:

$$\overrightarrow{w_J^{\text{new}}} = \beta(I \wedge \overrightarrow{w_J^{\text{old}}}) + (1-\beta)\overrightarrow{w_J^{\text{old}}}. \qquad \text{(A21)}$$

Fast learning can occur in fuzzy ART without causing catastrophic forgetting. Fast learning means that adaptive weights can reach their new equilibria on every learning trial. Our simulations were carried out under conditions of fast learning, for which $\beta = 1$ in (A21). Also, for each simulation, we used three fuzzy ART modules, one for each spatial scale of inputs I_i^g in (A13).

To make our simulations as efficient as possible, we implemented the following algorithmic instantiation of fuzzy ART at three spatial scales:

1. *Take the initial input from the current spatial scale g,* $\overrightarrow{I^g}_1$.
2. *Initialize the weight vector to the initial category node,* $\overrightarrow{w_1^g} = \overrightarrow{I_1^g}$ *, and set the node to be committed.*
3. *Present the next input* $\overrightarrow{I_n^g}$.
4. *Compute the activation to the category nodes via the signal function* (Carpenter & Gjaja, 1994), $T_j^g = \dfrac{|\overrightarrow{I_n^g} \wedge \overrightarrow{w_j^g}|}{\alpha + |\overrightarrow{w_j^g}|}$, *where M is the number of dimensions in the input vector, and* α *is the choice parameter.*
5. *Choose the category node* K^g *with the largest input; that is,* $T_{K^g}^g = \arg\max_j(\overrightarrow{T_j^g})$.
6. *Check to see if the chosen category satisfies vigilance:* $\dfrac{|\vec{I}_n^g \wedge \overrightarrow{w_J^g}|}{|\vec{I}_n^g|} \geq \rho$.
    a. *If vigilance is satisfied update the weights:* $\overrightarrow{w_{K^g}^{g,new}} = \beta(\overrightarrow{I_n^g} \wedge \overrightarrow{w_{K^g}^{g,old}}) + (1-\beta)\overrightarrow{w_{K^g}^{g,old}}$ *where* β *is the learning rate.*
    b. *If vigilance is not satisfied, then set* $T_J^g = -1$ *and go to Step 5. If all nodes have been exhausted (which would never happen in vivo), add a new category node, set the weights to* $\overrightarrow{I_n^g}$ *and set the node committed.*
7. *Go to Step 3 unless all inputs have been presented.*

Parameters. Model parameters were set as follows. For both simulations, $\alpha = .0001$ and $\beta = 1$. For the Zoccolan et al. (2007) data, $\rho = .1$. For the Akrami et al. (2009) data, $\rho = .9$.

Implementation Details. The model was implemented in Python using the NumPy (Dubois, Hinsen, & Hugunin, 1996), SciPy (Jones, Oliphant, Peterson, & others, 2001) and matplotlib (Hunter, 2007) libraries (sometimes referred to as PyLab). For possible extensions, parts of the model were also implemented in C using OpenCV.

Acknowledgements. This work was supported in part by CELEST, a National Science Foundation Science of Learning Center (SBE-0354379); and by the SyNAPSE program of the Defense Advanced Research Projects Agency (HR0011-09-C-0001).

References

Alvarez, G. A., & Cavanagh, P. (2008). Visual short-term memory operates more efficiently on boundary features than on surface features. *Perception and Psychophysics*, *70*(2), 346–364.

Banquet, J., & Grossberg, S. (1987). Probing cognitive processes through the structure of event-related potentials during learning: an experimental and theoretical analysis. *Applied Optics*, *26*(23), 4931-4946.

Bhatt, R., Carpenter, G. A., & Grossberg, S. (2007). Texture segregation by visual cortex: Perceptual grouping, attention, and learning. *Vision Research*, *47*(25), 3173–3211.

Bradski, G., & Grossberg, S. (1995). Fast-learning VIEWNET architectures for recognizing three-dimensional objects from multiple two-dimensional views. *Neural Networks*, *8*(7-8), 1053–1080.

Burt, P., & Adelson, E. (1983). The Laplacian Pyramid as a Compact Image Code. *Communications, IEEE Transactions on Communications*, *31*(4), 532–540.

Cao, Y., & Grossberg, S. (2005). A laminar cortical model of stereopsis and 3D surface perception: Closure and da Vinci stereopsis. *Spatial Vision*, *18*(5), 515–578.

Cao, Y., & Grossberg, S. (2011). Stereopsis and 3D surface perception by spiking neurons in laminar cortical circuits: A method for converting neural rate models into spiking models. Submitted for publication.

Cao, Y., Grossberg, S., & Markowitz, J. (2010). How does the brain rapidly learn and recorganize view- and positionally-invariant object representations in inferior temporal cortex? *Neural Network*s, submitted for publication.

Carpenter, G. A., & Gjaja, N. M. (1994). Fuzzy ART Choice Functions. *Proceedings of the World Congress on Neural Networks*, 713-722.

Carpenter, G. A., & Grossberg, S. (1990). ART3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. *Neural Networks*, *3*(2), 129–152.

Carpenter, G. A., & Grossberg, S. (1991). *Pattern Recognition by Self-Organizing Neural Networks*. Cambridge, MA, USA: MIT Press.

Carpenter, G. A., & Grossberg, S. (1993). Normal and amnesic learning, recognition and memory by a neural model of cortico-hippocampal interactions. *Trends in Neurosciences*, *16*(4), 131–137.

Carpenter, G. A., Grossberg, S., & Rosen, D. B. (1991). Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, *4*(6), 759–771.

Chiu, Y. C., & Yantis, S. (2009). A Domain-Independent Source of Cognitive Control for Task Sets: Shifting Spatial Attention and Switching Categorization Rules. *Journal of Neuroscience*, *29*(12), 3930.

Clay, R., & Alonso, J. M. (1995). Specificity of monosynaptic connections from thalamus to visual cortex. *Nature*, *378*(6554), 281–284.

Daniel, P. M., & Whitteridge, D. (1961). The representation of the visual field on the cerebral cortex in monkeys. *The Journal of Physiology*, *159*(2), 203.

Davidoff, J. (1991). *Cognition through color*. MIT Press Cambridge, MA.

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London. Series*

B: Biological Sciences, *353*(1373), 1245-1255.

DiCarlo, J. J., & Maunsell, J. H. (2003). Anterior Inferotemporal Neurons of Monkeys Engaged in Object Recognition Can be Highly Sensitive to Object Retinal Position. *Journal of Neurophysiology*, *89*(6), 3264–3278.

Dubois, P. F., Hinsen, K., & Hugunin, J. (1996). Numerical Python. *Computers in Physics*, *10*(3).

Elder, J. H., & Zucker, S. W. (1998). Evidence for boundary-specific grouping. *Vision Research*, *38*(1), 143–152.

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience*, *2*(10), 704–716.

Ettlinger, G., Iwai, E., Mishkin, M., & Rosvold, H. E. (1968). Visual discrimination in the monkey following serial ablation of inferotemporal and preoccipital cortex. *Journal of Comparative and Physiological Psychology*, *65*(1), 110–117.

Fang, L., & Grossberg, S. (2009). From stereogram to surface: How the brain sees the world in depth. *Spatial Vision*, *22*(1), 45–82.

Fazl, A., Grossberg, S., & Mingolla, E. (2009). View-invariant object category learning, recognition, and search: How spatial and object attention are coordinated using surface-based attentional shrouds. *Cognitive Psychology*, *58*(1), 1–48.

Fei, L. F., Fergus, R., & Perona, P. (2006). One-Shot Learning of Object Categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(4), 594–611.

Fischer, B. (1973). Overlap of receptive field centers and representation of the visual field in the cat's optic tract. *Vision Research*, *13*(11), 2113.

Foster, K., Gaska, J., Nagler, M., & Pollen, D. (1985). Spatial and temporal frequency selectivity of neurones in visual cortical areas V1 and V2 of the macaque monkey. *Journal of Physiology*, *365*(1), 331–363.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, *1*(2), 119–130.

Gochin, P. M. (1994). Properties of Simulated Neurons from a Model of Primate Inferior Temporal Cortex. *Cerebral Cortex*, *4*(5), 532–543.

Gross, C. G., Miranda, R. C. E., & Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex of the Macaque. *Journal of Neurophysiology*, *35*(1), 96–111.

Grossberg, S. (1973). Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, *52*(3), 213–257.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review*, *87*(1), 1-51.

Grossberg, S. (1994). 3-D vision and figure-ground separation by visual cortex. *Perception & Psychophysics*, *55*(1), 48-121.

Grossberg, S. (2003). How does the cerebral cortex work? Development, learning, attention, and 3-D vision by laminar circuits of visual cortex. *Behavioral and Cognitive Neuroscience Reviews*, *2*(1), 47–76.

Grossberg, S. (2007). Consciousness CLEARS the mind. *Neural Networks: The Official Journal of the International Neural Network Society*, *20*(9), 1040-1053. Grossberg, S., & Hong, S. (2006). A neural model of surface perception: lightness, anchoring, and filling-in. *Spatial Vision*, *19*(2-4), 263–321.

Grossberg, S., Kuhlmann, L., & Mingolla, E. (2007). A neural model of 3D shape-from-texture: Multiple-scale filtering, boundary grouping, and surface filling-in. *Vision Research*, *47*(5), 634–672.

Grossberg, S., & Merrill, J. W. (1996). The hippocampus and cerebellum in adaptively timed learning, recognition, and movement. *Journal of Cognitive Neuroscience*, *8*(3), 257–277.

Grossberg, S., & Mingolla, E. (1985). Neural dynamics of perceptual grouping: textures, boundaries, and emergent segmentations. *Perceptual Psychophysics*, *38*(2), 141–71.

Grossberg, S., & Raizada, R. D. S. (2000). Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. *Vision Research*, *40*(10-12), 1413–1432.

Grossberg, S., & Seidman, D. (2006). Neural dynamics of autistic behaviors: cognitive, emotional, and timing substrates. *Psychological Review*, *113*(3), 483–525.

Grossberg, S., & Versace, M. (2008). Spikes, synchrony, and attentive learning by laminar thalamocortical circuits. *Brain Research*, *1218*, 278–312.

Grossberg, S., & Williamson, J. R. (1999). A self-organizing neural system for learning to recognize textured scenes. *Vision Research*, *39*(7), 1385–1406.

Grossberg, S., & Yazdanbakhsh, A. (2005). Laminar cortical dynamics of 3D surface perception: stratification, transparency, and neon color spreading. *Vision Research*, *45*(13), 1725–1743.

Hirsch, J. A., Alonso, J. M., Reid, C. R., & Martinez, L. M. (1998). Synaptic Integration in Striate Cortical Simple Cells. *Journal of Neuroscience*, *18*(22), 9517–9528.

Hodgkin, A. L. (1964). *The conduction of the nervous impulse*. Springfield, IL: Charles C. Thomas.

Horton, J. C., & Hoyt, W. F. (1991). The representation of the visual field in human striate cortex. A revision of the classic Holmes map. *Archives of Ophthalmology*, *109*(6), 816-824.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*, 574–591.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*, 106–154.

Hunter, J. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90-95.

Jones, E., Oliphant, T., Peterson, P., & others. (2001). *SciPy: Open source scientific tools for Python*. Retrieved from http://www.scipy.org/

Julesz, B., & Schumer, A. (1981). Early visual perception. *Annual Review of Psychology*, *32*, 575-627.

Kelly, F., & Grossberg, S. (2000). Neural dynamics of 3-D surface perception: Figure-ground separation and lightness perception. *Perception and Psychophysics*, *62*(8), 1596–1618.

Lamme, V. A., Rodriguez-Rodriguez, V., & Spekreijse, H. (1999). Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the macaque monkey. *Cerebral Cortex*, *9*(4), 406.

Li, L., Miller, E. K., & Desimone, R. (1993). The representation of stimulus familiarity in anterior inferior temporal cortex. *Journal of Neurophysiology*, *69*(6), 1918–1929.

Mingolla, E., Ross, W., & Grossberg, S. (1999). A neural network for enhancing boundaries and surfaces in synthetic aperture radar images. *Neural Networks*, *12*(3), 499–511.

Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978). Spatial and temporal contrast sensitivity of neurones in areas 17 and 18 of the cat's visual cortex. *Journal of Physiology*, *283*(1), 101–120.

Pollen, D. A. (1999). On the neural correlates of visual perception. *Cerebral Cortex*, *9*(1), 4.

Porat, M., & Zeevi, Y. Y. (1988). The generalized Gabor scheme of image representation in biological and machine vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *10*(4), 452–468.

Raizada, R. D., & Grossberg, S. (2003). Towards a theory of the laminar architecture of cerebral cortex: Computational clues from the visual system. *Cerebral Cortex*, *13*(1), 100.

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*(2), 168–185.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, *12*(2), 162–168.

Rogers-Ramachandran, D. C., & Ramachandran, V. S. (1998). Psychophysical evidence for boundary and surface systems in human vision. *Vision Research*, *38*(1), 71–77.

Rosenfeld, A., & Thurston, M. (1971). Edge and Curve Detection for Visual Scene Analysis. *IEEE Transactions on Computers*, *C-20*(5), 562–569.

Schwartz, E. L. (1980). Computational anatomy and functional architecture of striate cortex: A spatial mapping approach to perceptual coding. *Vision Research*, *20*(8), 645–669.

Schwartz, E. L., Desimone, R., Albright, T. D., & Gross, C. G. (1983). Shape Recognition and Inferior Temporal Neurons. *Proceedings of the National Academy of Sciences*, *80*(18), 5776–5778.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007). A quantitative theory of immediate visual recognition. *Progress in Brain Research*, *165*, 33–56.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, *104*(15).

Stringer, S. M., Perry, G., Rolls, E. T., & Proske, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biological Cybernetics*, *94*(2), 128–142.

Tootell, R. B., Silverman, M. S., Switkes, E., & De Valois, R. L. (1982). Deoxyglucose analysis of retinotopic organization in primate striate cortex. *Science*, *218*(4575), 902.

Tyler, C. W. (1975). Spatial organization of binocular disparity sensitivity. *Vision Research*, *15*(5), 583-590.

Van Essen, D. C., Newsome, W. T., & Maunsell, J. H. (1984). The visual field representation in striate cortex of the macaque monkey: asymmetries, anisotropies, and individual variability. *Vision Research*, *24*(5), 429-448.

von der Heydt, R., & Peterhans, E. (1989). Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *Journal of Neuroscience*, *9*(5), 1731.

von der Heydt, R., Peterhans, E., & Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science*, *224*(4654), 1260–1262.

Watson, A. B. (1987). The cortex transform: rapid computation of simulated neural images. *Computer Vision, Graphics, and Image Processing*, *39*(3), 311–327.

Zoccolan, D., Kouh, M., Poggio, T., & DiCarlo, J. J. (2007). Trade-Off between Object Selectivity and Tolerance in Monkey Inferotemporal Cortex. *Journal of Neuroscience*, *27*(45), 12292-12307.