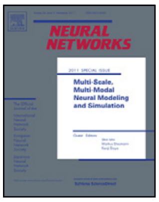
Neural Networks 24 (2011) 1050–1061

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



# How does the brain rapidly learn and reorganize view-invariant and position-invariant object representations in the inferotemporal cortex?

Yongqiang Cao, Stephen Grossberg*, Jeffrey Markowitz

*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*

ARTICLE INFO

*Article history:*
Received 16 December 2010
Received in revised form 10 April 2011
Accepted 12 April 2011

*Keywords:*
Inferotemporal cortex
Category learning
Invariant object recognition
View category
Persistent activity
Spatial attention
Disengage attention
Sustained attention
Transient attention
Cortical magnification factor
Target swapping
Adaptive resonance theory

ABSTRACT

All primates depend for their survival on being able to rapidly learn about and recognize objects. Objects may be visually detected at multiple positions, sizes, and viewpoints. How does the brain rapidly learn and recognize objects while scanning a scene with eye movements, without causing a combinatorial explosion in the number of cells that are needed? How does the brain avoid the problem of erroneously classifying parts of different objects together at the same or different positions in a visual scene? In monkeys and humans, a key area for such invariant object category learning and recognition is the inferotemporal cortex (IT). A neural model is proposed to explain how spatial and object attention coordinate the ability of IT to learn invariant category representations of objects that are seen at multiple positions, sizes, and viewpoints. The model clarifies how interactions within a hierarchy of processing stages in the visual brain accomplish this. These stages include the retina, lateral geniculate nucleus, and cortical areas V1, V2, V4, and IT in the brain's What cortical stream, as they interact with spatial attention processes within the parietal cortex of the Where cortical stream. The model builds upon the ARTSCAN model, which proposed how view-invariant object representations are generated. The positional ARTSCAN (pARTSCAN) model proposes how the following additional processes in the What cortical processing stream also enable position-invariant object representations to be learned: IT cells with persistent activity, and a combination of normalizing object category competition and a view-to-object learning law which together ensure that unambiguous views have a larger effect on object recognition than ambiguous views. The model explains how such invariant learning can be fooled when monkeys, or other primates, are presented with an object that is swapped with another object during eye movements to foveate the original object. The swapping procedure is predicted to prevent the reset of spatial attention, which would otherwise keep the representations of multiple objects from being combined by learning. Li and DiCarlo (2008) have presented neurophysiological data from monkeys showing how unsupervised natural experience in a target swapping experiment can rapidly alter object representations in IT. The model quantitatively simulates the swapping data by showing how the swapping procedure fools the spatial attention mechanism. More generally, the model provides a unifying framework, and testable predictions in both monkeys and humans, for understanding object learning data using neurophysiological methods in monkeys, and spatial attention, episodic learning, and memory retrieval data using functional imaging methods in humans.

## 1. Introduction

The brain effortlessly learns to recognize objects that are seen at multiple positions, sizes, and viewpoints. How does the brain rapidly learn to recognize objects while scanning a scene with eye movements, without causing a combinatorial explosion in the number of cells that are needed? How does the brain avoid the problem of erroneously classifying parts of different objects together? In monkeys and humans, a key area for such invariant object learning and recognition is the inferotemporal cortex (IT). A neural model is proposed to explain how spatial and object attention coordinate the ability of IT to learn representations of object categories that are seen at multiple positions, sizes, and viewpoints. Such invariant object category learning and recognition can be achieved using interactions between a hierarchy of processing stages in the visual brain. These stages include the retina, lateral geniculate nucleus, and cortical areas V1, V2, V4, and IT in the brain's What cortical stream, as

* Corresponding address: Center for Adaptive Systems, Department of Cognitive and Neural Systems, Boston University, 677 Beacon Street, Boston, MA, 02215, USA. Tel.: +1 617 353 7858, 1 617 353 7857; fax: +1 617 353 7755.
*E-mail address:* steve@bu.edu (S. Grossberg).

0893-6080/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2011.04.004

they interact with spatial attention processes within the parietal cortex of the Where cortical stream. The model builds upon the ARTSCAN model (Fazl, Grossberg, & Mingolla, 2009; Grossberg, 2009), which proposed how view-invariant object representations may be learned and recognized.

A key prediction of the ARTSCAN model is how the reset of spatial attention in the Where cortical stream prevents views of different objects from being learned as part of the same invariant IT category. The positional ARTSCAN (pARTSCAN) model that is developed in the current article proposes how the following additional processes in the What cortical processing stream also enable position-invariant object representations to be learned: IT cells with persistent activity, and a combination of normalizing object category competition and a view-to-object learning law which together ensure that unambiguous views have a larger effect on object recognition than ambiguous views. The model is tested by simulating neurophysiological data from a target swapping experiment of Li and DiCarlo (2008) that is predicted to fool the spatial attentional reset mechanisms which usually keep different object views separated during learning.

Many electrophysiological experiments have shown that cells in the inferotemporal (IT) cortex respond to the same object at different retinal positions; for example, many IT cells show little attenuation in firing rate across object translations (Booth & Rolls, 1998; Desimone & Gross, 1979; Gross, Rocha-Miranda, & Bender, 1972; Ito, Tamura, Fujita, & Tanaka, 1995; Schwartz, Desimone, Albright, & Gross, 1983). The target swapping experiment of Li and DiCarlo (2008) showed, in addition, how the positional selectivity of cells in IT can be altered by experience. Their experiment was divided into two exposure phases, in which two extra-foveal positions (3° above or below the center of gaze) were prechosen as swap and non-swap positions. The experiment studied IT neuronal responses to two objects that initially elicited strong (object $P$, preferred) and moderate (object $N$, non-preferred) responses at the two positions. The monkey always began a learning trial looking at a fixation point. During a "normal exposure", when an object appeared at the prechosen non-swap position, the monkey quickly moved its eyes to it with a saccadic eye movement that brought its image to the fovea. During a "swap exposure", in which an object appeared at the prechosen swap position, the object $P$ (or $N$) was always swapped for the other object $N$ (or $P$) during the saccade. Li and DiCarlo found that IT neuron selectivity to objects $P$ and $N$ at the swap position was reversed with increasing exposure (see Fig. 1(A)), but there was little or no change at the non-swap position.

The pARTSCAN model (Fig. 2) quantitatively explains and simulates the Li and DiCarlo data as a manifestation of the mechanisms whereby the brain learns position-invariant object representations. Some prominent efforts to model IT have built invariant representations using a hierarchy of feedforward filters leading to a learned category choice (Bradski & Grossberg, 1995; Grossberg & Huang, 2009; Riesenhuber & Poggio, 1999, 2000, 2002), or through grouping object translations through time (Fazl et al., 2009; Wallis & Rolls, 1997). The pARTSCAN model proposes how the brain learns position-invariant object representations that are consistent with the Li and DiCarlo swapping data. In particular, the pARTSCAN model, as in the ARTSCAN model on which it builds, proposes how multiple brain processing stages, beginning in the retina and lateral geniculate nucleus (LGN), and proceeding through cortical areas V1, V2, V4, and IT in the What cortical stream, can gradually learn such position-invariant object representations, as they interact with Where cortical processes stages in the parietal cortex.

The ARTSCAN model proposes how an object's surface representation in cortical area V4 generates a form-fitting distribution of spatial attention, or "attentional shroud", in the parietal cortex of the Where cortical stream. All surface representations dynamically compete for spatial attention to form a shroud. The winning shroud (or shrouds; see Foley, Grossberg, and Mingolla (submitted for publication) for simulations of multifocal attention) remains active due to a surface-shroud resonance that persists during active scanning of the object with eye movements. The active shroud regulates eye movements and category learning about the attended object in the following way.

The first view-specific category to be learned for the attended object also activates a cell population at a higher processing stage. This cell population will become a view-invariant object category. Both types of category are assumed to form in the IT cortex of the What cortical stream. As the eyes explore different views of the object, previously active view-specific categories are reset to enable new view-specific categories to be learned. What prevents the emerging view-invariant object category from also being reset? The shroud maintains the activity of the emerging view-invariant category representation by inhibiting a reset mechanism, also predicted to be in the parietal cortex, that would otherwise inhibit the view-invariant category. As a result, all the view-specific categories can be linked through associative learning to the emerging view-invariant object category. Indeed, these associative linkages create the view invariance property.

Shroud collapse disinhibits the reset signal, which in turn inhibits the active view-invariant category. Then a new shroud, corresponding to a different object, forms in the Where cortical stream as new view-specific and view-invariant categories of the new object are learned in the What cortical stream. The model hereby mechanistically clarifies basic properties of spatial attention shifts (engage, move, disengage) and inhibition of return. As noted in Section 4, the concepts of shroud persistence and reset clarify traditional ideas about sustained and transient attention, respectively.

The ARTSCAN model does not, however, explain how position-invariant object categories are learned and recognized. The current article proposes what additional brain mechanisms are needed to learn position-invariant object categories. These new mechanisms include a new functional role for cells with persistent activity in IT (see Brunel, 2003; Fuster & Jervey, 1981; Miyashita & Chang, 1988; Tomita, Ohbayashi, Nakahara, Hasegawa, & Miyashita, 1999) and a competitive learning law whereby more predictive unambiguous object views learn to have a larger effect on object recognition than less predictive ambiguous views.

The pARTSCAN model quantitatively simulates the swapping data by showing how the swapping procedure fools the spatial attentional shroud mechanism that usually is reset when a new object is presented, thereby preventing multiple objects from learning to activate the same invariant object category. The model predicts that the shroud of the previous object is not reset during the swap with another object. Persistence of this attentional shroud across swaps leads to rapid reshaping of IT receptive fields through unsupervised natural visual experience when it interacts with IT persistent activity and competitive learning. In addition to these prediction, which can be tested in monkeys, a prediction is made in Section 4 about how to test the shroud hypothesis during a swapping experiment using fMRI in humans. The same combination of brain mechanisms can also explain how swapping targets of different sizes can lead to rapid learning of the corresponding mixtures of object views at different sizes (Li & DiCarlo, 2010).

## 2. Results

### 2.1. Model processing stages

The model consists of the following processing stages. See Fig. 2. These stages are described heuristically in this section and mathematically in Section 5.

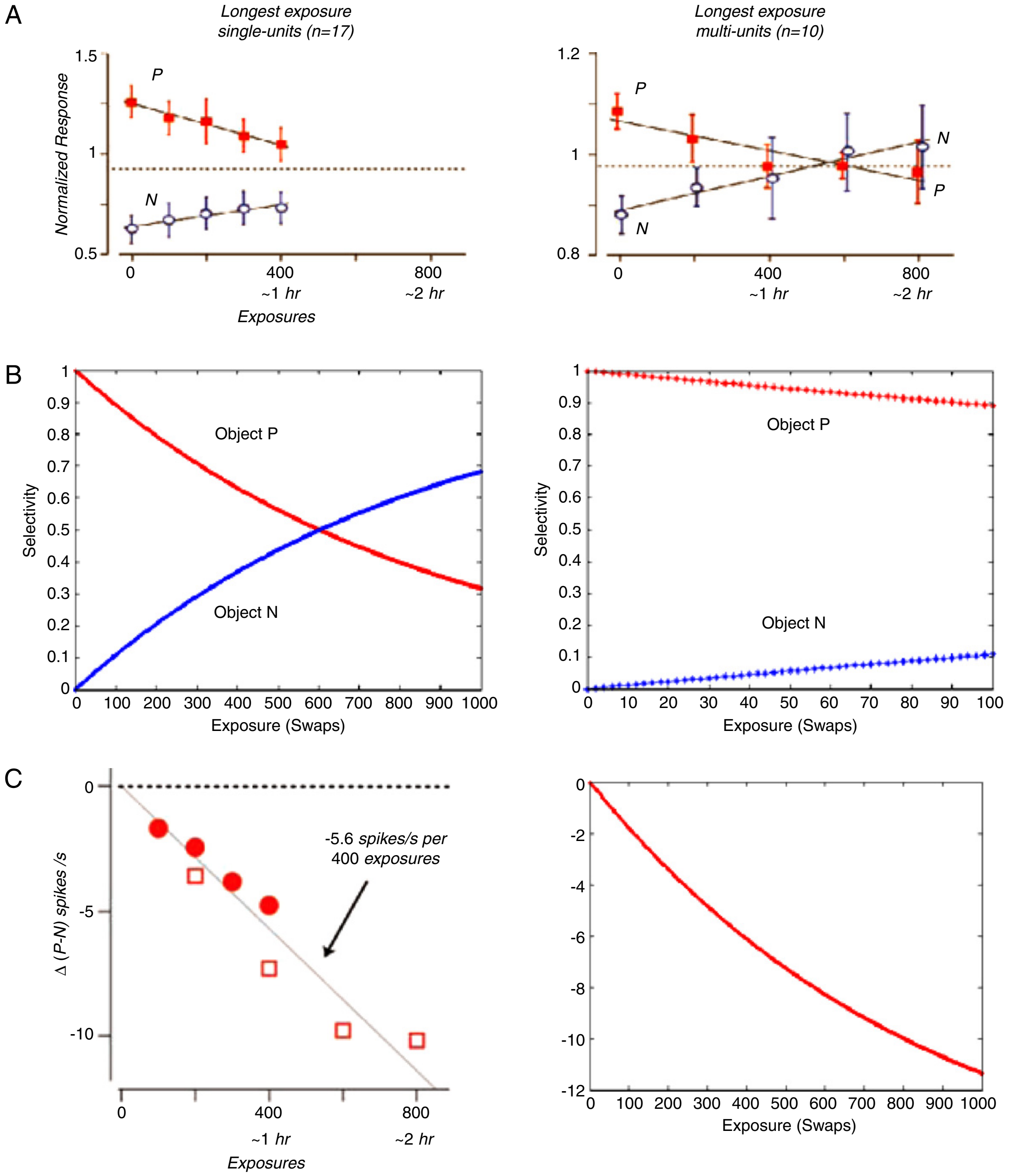


**Fig. 1.** Experimental data and model simulation. (A) The Li and DiCarlo experimental data, which shows that IT neuron selectivity to objects $P$ and $N$ at swap position is reversed with increasing swap exposures. (B) Left: The selectivity of model IT object category neuron $O_p$ to views of objects $P$ and $N$ at the swap position as a function of swap exposures. These selectivity values are computed as the normalized object category neuron activities that are defined by Eq. (21). They have the same sizes as the view-to-object connection weights that are defined by Eq. (22). The selectivity to objects $P$ and $N$ reverses at about 600 swaps, which is consistent with the experimental data in A. Increasing the learning rate in Eq. (24) will advance the reversal time, and vice versa. Right: A zoom for the first 100 swaps of the model IT neuron. (C) Left: Mean object selectivity change as a function of the number of swap exposures for experimental IT neurons, $\Delta(P-N) = (P-N)_{\text{post-exposure}} - (P-N)_{\text{pre-exposure}}$. Right: Normalized object selectivity change as a function of the number of swap exposures for the model IT object neuron. The normalized object selectivity change is computed by $\Delta(W_P - W_N)C$, where $W_P$ and $W_N$ are the connection weights of the model IT object neuron with view integrator neurons of objects $P$ and $N$, and $C$ is a normalization constant (8.3).
*Source:* (A) and (C) Left are adapted with permission from Li and DiCarlo (2008).

*Contrast normalization and discounting the illuminant.* The contrasts in each input image are normalized, and background illumination is discounted, in the simplified model retina and LGN by an on-center off-surround network whose cells obey membrane, or shunting, equations (Grossberg & Todorovic, 1988; Werblin, 1971). This network defines ON cells that are turned on by image luminance increments. A parallel off-center on-surround network defines OFF cells that are turned on by image luminance decrements.

*Cortical magnification.* The foveal area is over-represented and the extra-foveal area under-represented due to the cortical magnification factor whereby outputs from the LGN are represented in cortical area V1 (Daniel & Whitteridge, 1961; Fischer, 1973; Horton & Hoyt, 1991; Schwartz, 1980; Tootell, Silverman, Switkes, & DeValois, 1982; Van Essen, Newsome, & Maunsell, 1984). Cortical magnification is carried out by a log-polar transform (Schwartz, 1980).

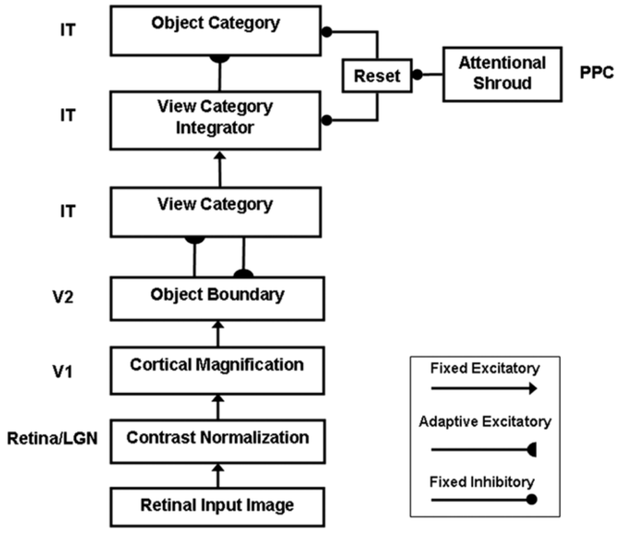


**Fig. 2.** Model processing stages. See the text for details.

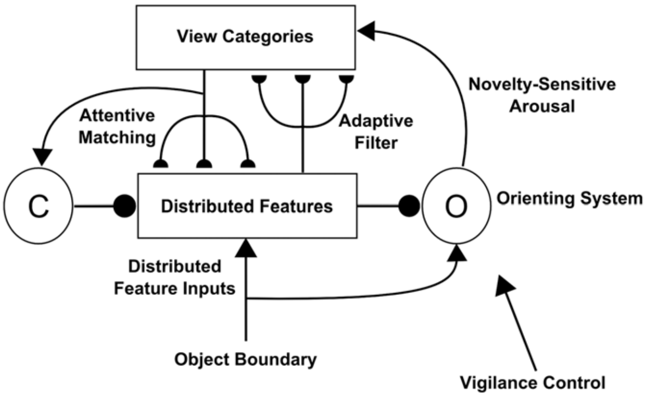


**Fig. 3.** View category learning circuit. Top-down expectations are attentively matched against bottom-up input features. Attended critical features are learned. A big enough mismatch (i.e., one that does not satisfy the vigilance criterion in the orienting system $O$) causes a burst of novelty-sensitive arousal from $O$ that inhibits the active category and triggers search for and learning of a better matching category. $C$: biased competition; black arrowheads: excitatory connections; black disks: inhibitory connections; black semicircles: learned connections.

*Object boundaries*. The contrast-normalized and log-polar-transformed ON and OFF cell outputs are added at each position to form object boundary representations in the model cortical area V2 (Peterhans & von der Heydt, 1989; von der Heydt & Peterhans, 1989; von der Heydt, Peterhans, & Baumgartner, 1984). Many psychophysical studies have supported the prediction that boundaries and surfaces are the perceptual units of three-dimensional (3D) vision (Grossberg, 1987). Boundaries are often sufficient to enable object recognition (Alvarez & Cavanagh, 2008; Bradski & Grossberg, 1995; Davidoff, 1991; Elder & Zucker, 1998; Grossberg, 1994; Grossberg & Mingolla, 1985; Lamme, Rodriguez-Rodriguez, & Spekreijse, 1999; Peterhans & von der Heydt, 1989; Rogers-Ramachandran & Ramachandran, 1998), and that is the case for the targeted neurophysiological data.

To quantitatively fit these data, no further boundary processing beyond contrast-normalized ON and OFF cell outputs is needed. However, the current simplified model is consistent with the more sophisticated 3D LAMNART model of how depth-selective boundaries may be formed, completed, and used to separate figures from their backgrounds. The current model can consistently be extended to include these additional mechanisms in cases where more ambiguous natural images require further processing; see, e.g., Cao and Grossberg (2005), Fang and Grossberg (2009) and Grossberg and Yazdanbakhsh (2005).

*View category learning via top-down attentive matching and memory search*. Accumulating evidence supports the hypothesis that the inferotemporal (IT) cortex learns to categorize objects. Some IT cells learn about individual views of an object. These neurons learn to categorize a seemingly infinite set of views into finitely many view-specific categories (Bradski & Grossberg, 1995; Bulthoff & Edelman, 1992; Bulthoff, Edelman, & Tarr, 1995; Carpenter & Ross, 1993; Fazl et al., 2009; Logothetis, Pauls, Bulthoff, & Poggio, 1994; Poggio & Edelman, 1990; Riesenhuber & Poggio, 2000; Seibert & Waxman, 1992). Each view-specific category can learn to respond to modest changes in object boundaries due to different orientations, sizes, and viewpoints. However, if any of these factors changes too much as the eyes move on an object surface, or the surface moves relative to the eyes, then an active view-specific category is reset. As this happens through time, view-specific category neurons that respond to different views of the same object learn to activate the same neuronal population, creating a view-invariant category representation of the object.

In the pARTSCAN model, contrast-normalized, log-polar transformed boundaries are the inputs to the model IT, which rapidly learns view categories from these individual experiences through time. In order to ensure the stability of this learning though time, in addition to bottom-up learning within the adaptive filters that activate IT view categories, the activated view categories read out learned top-down expectations whose prototypes are matched against the boundary feature patterns (Fig. 3; Bhatt, Carpenter, & Grossberg, 2007; Carpenter & Grossberg, 1987, 1991, 1993; Grossberg, 1980, 2007; Miller, Li, & Desimone, 1993). This matching process focuses object attention on the pattern of critical boundary features that have been learned by the prototype of the expectation. This process of focusing attention is realized by a top-down, modulatory on-center, driving off-surround network (Bhatt et al., 2007; Carpenter & Grossberg, 1987, 1991), which explains and has 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 causing catastrophic forgetting. Adaptive resonance theory, or ART, predicts that this dilemma is solved in the following way. If a top-down match is good enough, it can cause a synchronous *resonance* between the attended critical feature pattern and the selected category via bottom-up and top-down signal exchanges. Such a resonance triggers fast learning in the adaptive weights, or long-term memory traces, that occur in the bottom-up and top-down pathways that carry the resonant 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, as it simultaneously dynamically buffers system memories against catastrophic forgetting. If the match is not good enough, then a mismatch occurs between the learned top-down expectation and the bottom-up featural inputs. Such a mismatch triggers a memory search, or bout of hypothesis testing, that ends in selection and learning of a better-matching category.

An *orienting system* regulates memory search by computing the degree of match between the bottom-up input pattern and the learned top-down expectation (Fig. 3). 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 (Otto & Eichenbaum, 1992; Vinogradova, 1975). The criterion for a good enough match can depend upon the task (Spitzer, Desimone, & Moran, 1998). A task-selective *vigilance* parameter in the model determines how strict the matching criterion is, with higher vigilance requiring a better match to trigger resonance and learning (Carpenter & Grossberg, 1987, 1991, 1993). 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), whereas low vigilance leads to the learning of more general and abstract categories (e.g., a general face category).

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, which 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, resonance, and vigilance mechanisms to learn cortical recognition codes, see Carpenter and Grossberg (1991, 1993), Engel, Fries, and Singer (2001), Grossberg (2003, 2007), Pollen (1999) and Raizada and Grossberg (2003).

In a detailed laminar model of thalamocortical interactions, Grossberg and Versace (2008) have predicted how vigilance may be controlled by novelty-sensitive corticothalamic mismatches that activate the nonspecific thalamus, which in turn activates acetylcholine release via the nucleus basalis of Meynert, thereby increasing excitability of cortical layer 5 cells, and leading to reset in layer 4 via layer 5-to-6-to-4 signals. The inability to flexibly alter vigilance to match task demands has been predicted to occur in certain mental disorders, such as autism (Grossberg & Seidman, 2006).

*View category integrator and persistent activity*. One important new feature of our model is view category integrator neurons, whose persistent activities are predicted to play a key role in learning position-invariant object representations. View category integrator neurons receive input from view-specific category neurons (Fig. 2). They preserve activity of view category neurons that are reset as new object views are seen due to either object motion or eye movements. The properties of these model neurons are consistent with neurophysiological data about cells with persistent activity in IT (Brunel, 2003; Fuster & Jervey, 1981; Miyashita & Chang, 1988; Tomita et al., 1999). This persistent activity is modeled by a leaky integrator with a slow decay rate.

What functional role in position-invariant object category learning do view category integrator neurons play? Suppose that view category integrator neurons do not exist. As shown in Fig. 4(A), when an object $P$ forms an image in the fovea, the brain begins to learn a view-specific category $V$, which then selects a cell population that will become, as multiple views are associated with it, a position-invariant and view-invariant object category $O$. In Fig. 4(B), the next time when $P$ appears in an extra-foveal position, the brain begins to learn a new view-specific category $V1$, which activates object category $O1$. A saccade then brings $P$'s image to the fovea (Fig. 4(C)), and activates the previously learned view-specific category $V$ and object category $O$. Since view category $V1$ is shut off with the saccade, it cannot learn to be associated with object category $O$. Its learned association to $O1$ thus persists. As a result, object $P$ learns two object categories $O$ and $O1$ for the two retinal positions. The same reasoning holds when $P$ appears at any new extra-foveal position in Fig. 4(B), so $P$ will learn a different object category for each extra-foveal position. Therefore, no position-invariant object category is learned for $P$.

View category integrator neurons help to prevent an object from creating multiple object categories at different positions. In Fig. 4(D) and (E), view category integrator neurons $T$ and $T1$ preserve the activities of view categories $V$ and $V1$ after a saccade occurs. In Fig. 4(F), although $V1$ is shut off with the saccade, $T1$ is still active due to persistent activity. $T1$ can thus be associated through learning with object category $O$. The same is true for $P$ appearing at any other extra-foveal position. As a result, the brain can now learn a position-invariant object category $O$ for $P$.

*Regulation of object category learning by spatial attentional resonance and reset*. As shown in Fig. 4(F), object category neurons receive inputs from view category integrator neurons. When an object is explored during eye movements, the eyes quickly move among the salient features on the object surface (Yarbus, 1967). This exploration leads to the learning of multiple view-specific categories.

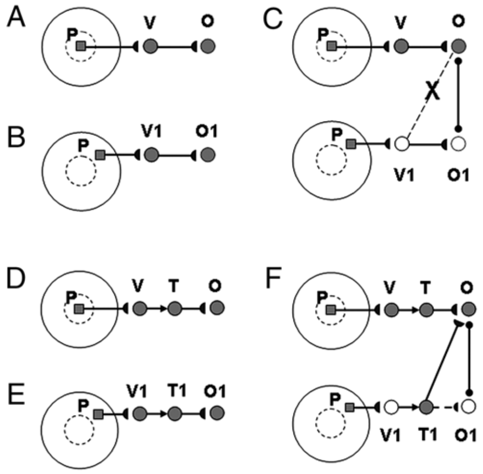


**Fig. 4.** How view category integrator neurons help to learn position-invariant object categories. Large circles denote retinas; dotted circles denote foveas within the retinas; filled little squares denote object retinal images; filled little circles denote active neurons; open little circles denote inactive neurons; lines with black disks denote inhibitory connections; black semidisc denote learned connections; black arrowheads denote excitatory unlearned connections; the dotted line with $X$ denotes that no learned connection can occur; the dotted line with semidisc denotes losing learned connection. See the text for details.

Although view category neurons are reset with each eye movement, an object category neuron needs to remain active until the eyes move to examine a different object. When the eyes do move to inspect a different object, then the previously active object category neurons need to be reset, so that they are not erroneously associated with view categories from a different object. In this way, multiple learned view-specific categories for the same object can be associated with one view-invariant object category through learning.

As outlined in Section 1, the ARTSCAN model of Fazl et al. (2009) proposes how the brain prevents an emerging view-invariant category from being reset, even while its individual view categories are reset, by using an attentional signal from the parietal cortex. This attentional signal arises when an object's surface representation, say in cortical area $V4$, resonates with a form-fitting distribution of spatial attention in the parietal cortex. Such a form-fitting distribution of spatial attention is called an *attentional shroud* (Tyler & Kontsevich, 1995). The ARTSCAN model explains how a shroud remains active while the observer attends the object, even while the observer's eye movements actively explore the object to inspect new views. An active shroud inhibits reset of an object category (Fig. 2), thereby enabling the object category to be associated with multiple learned view categories of the same object. In this way, the object category becomes a view-invariant category. When the eyes move to a different object, the previously active shroud is shut off and the corresponding object category neuron is inhibited, or reset, to enable a new object to be attended and its object category to be learned.

When a view-specific category focuses attention on its salient boundary features (Fig. 2), this is accomplished by object attention in the What cortical stream. Thus, the ARTSCAN model proposes how *spatial attention* in the Where cortical processing stream regulates *object attention* in the What cortical processing stream.

The coordinated interaction between spatial attention and object category learning and attention prevents view categories that correspond to different objects from mistakenly being associated with the same object category during unsupervised scanning and learning about the world. The ARTSCAN prediction that a spatial attention shift (shroud collapse) causes a transient reset burst in the parietal cortex that, in turn, causes a shift in categorization rules (new object category activation) has been supported by experiments using rapid event-related fMRI (Chiu & Yantis, 2009). These and related results (e.g., Cabeza, Ciaramelli, Olson, & Moscovitch, 2008; Corbetta, Kincade, Ollinger, McAvoy, & Shulman, 2000; Yantis et al., 2002) suggest that different regions of the parietal cortex maintain sustained attention to a currently attended object (shroud) and control transient attention switching (reset burst) to a different object.

When a new object is surreptitiously swapped with a previous item during a saccadic eye movement, the shroud reset mechanism is fooled and does not reset. This property helps to explain how more than one object can learn to activate the same invariant object category. The ARTSCAN model hereby predicts that swapping views of different objects at the same position can rapidly reshape invariant object representations. However, ARTSCAN cannot explain the Li and DiCarlo (2008) swapping experiment, because this experiment also involves changes of object position across space.

In order to explain how object categories can learn to become position-invariant as well as view-invariant, the pARTSCAN model further predicts that view category integrator neurons are interpolated between view category neurons and object category neurons, and that the view integrator neurons are reset when a shift of spatial attention occurs—that is, when an attentional shroud collapses—at the same time that object category neurons are reset (Fig. 2). Then the effects of swaps that exchange object positions, sizes, and views can all be explained by the same combination of model mechanisms.

A complete simulation of the dynamics of shroud formation and reset is given in Fazl et al. (2009). For simplicity, the current model assumes that the shroud is reset at the appropriate times.

*Ambiguous views and learning of object categories*. When looked at from certain viewpoints, different objects can sometimes form the same or very similar images in the retina. This causes an *ambiguous view problem*: a single view category can be associated with more than one object category. An object with ambiguous views can be identified by adding distinguishable views through active exploration with eye movements.

In the pARTSCAN model, view category integrator neurons activate object category neurons via signals that are gated by learned weights, or long-term memory traces (Fig. 2). The active view category integrator neurons learn to strengthen their connection weights to the winning object category neuron, while weakening their connection weights with losing object category neurons. This is achieved in the model by a combination of normalizing competition among the object category neurons, and a view-to-object learning law called the *outstar learning rule* (Grossberg, 1968, 1980) that together enable distinguishable views to learn stronger learned connections with associated object category neurons than ambiguous views (see the Ambiguous View Learning Theorem in the Appendix). This sort of normalized learning, when embedded within the model circuitry, is also sufficient to simulate the Li and DiCarlo data property that an IT neuron which is initially strongly selective to object $P$ loses its selectivity to $P$ with increasing swaps at the swap position (Fig. 1). Without these model mechanisms, the IT neuron's selectivity to $P$ would remain unchanged even at the swap position. See Section 3 for more details. This learning law differs from the *instar learning* law, variants of which are used to learn self-organizing maps and view-specific categories (Carpenter & Grossberg, 1987, 1991; Fazl et al., 2009; Grossberg, 1976, 1980; Kohonen, 1989).

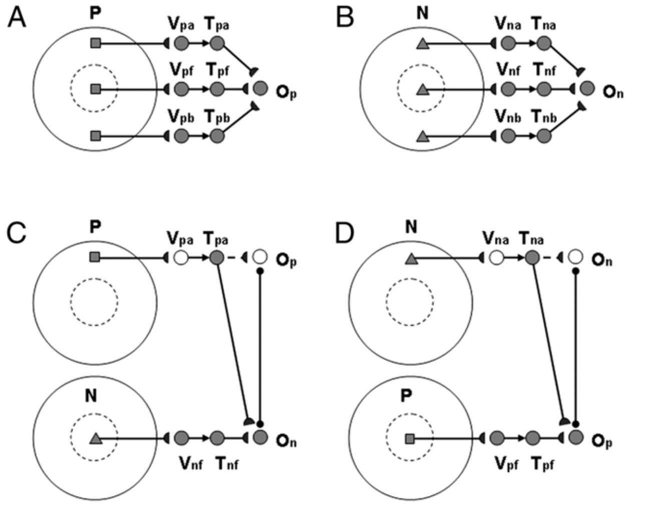


**Fig. 5.** How the model explains the Li and DiCarlo data. Filled little squares: object $P$'s retinal images; filled little triangles: object $N$'s retinal images. The other symbols are the same as in Fig. 4. See the text for details.

## 3. Model simulations

The swapping simulations used the two objects, boat ("$P$") and bowl ("$N$"), as stimuli that were used in the Li and DiCarlo experiment. In the simulated retina, an above-foveal position was predefined as the swap position and a below-foveal position was predefined as the non-swap position. In order to simulate normal daily experience, the two objects were first learned at three positions (the predefined swap position, non-swap position, and fovea) by 10,000 normal exposures. This experience led to learning of two object categories, each associated with three learned view categories via view category integrators (Fig. 5(A) and (B)). Next, following the procedure described in the Li and DiCarlo experiment, the swap exposure phase was simulated with 1000 swaps at the swap position, which was equally divided into 500 $P$-to-$N$ swaps and 500 $N$-to-$P$ swaps. The two types of swaps were presented alternately.

Each time, when object $P$ ("boat") appeared at the swap position (Fig. 5(C)), it activated prelearned object category $O_p$ through view category $V_{pa}$ and associated view category integrator $T_{pa}$. During a saccade, $P$ was replaced by object $N$. This brought $N$'s image to the fovea, and activated the prelearned view category $V_{nf}$, associated view category integrator $T_{nf}$, and object category $O_n$ of object $N$. Since $O_n$ receives input from the fovea, it wins the competition with $O_p$. Although view category $V_{pa}$ shuts off with the saccade, view category integrator $T_{pa}$ stays active due to its persistent activity. The view category integrator $T_{pa}$ can thus strengthen its learned connection with the winning object category $O_n$, and weaken its connection with the losing object category $O_p$. As a result, the selectivity, or learned weight, of object category neuron $O_p$ for object $P$ at the swap position decreases.

Similarly, for every $N$-to-$P$ swap where object $N$ ("bowl") appeared at the above-fovea swap position (Fig. 5(D)), the view category integrator $T_{na}$ learns to strengthen its connection with the winning object category $O_p$, and weaken its connection with the losing object category $O_n$. The selectivity (or learned weight) of object category neuron $O_p$ for object $N$ at the swap position hereby increases. As a result, the selectivity of object category neuron $O_p$ for objects $P$ and $N$ at the swap position is gradually reversed. Fig. 1(B) shows how the selectivity of neuron $O_p$ at the swap position reverses with swaps. This is consistent with the experimental data (Fig. 1(A)), where the reversal occurs at about 600 swaps. Fig. 1(C) shows how object selectivity changes as a function of the number of swap exposures in both experimental

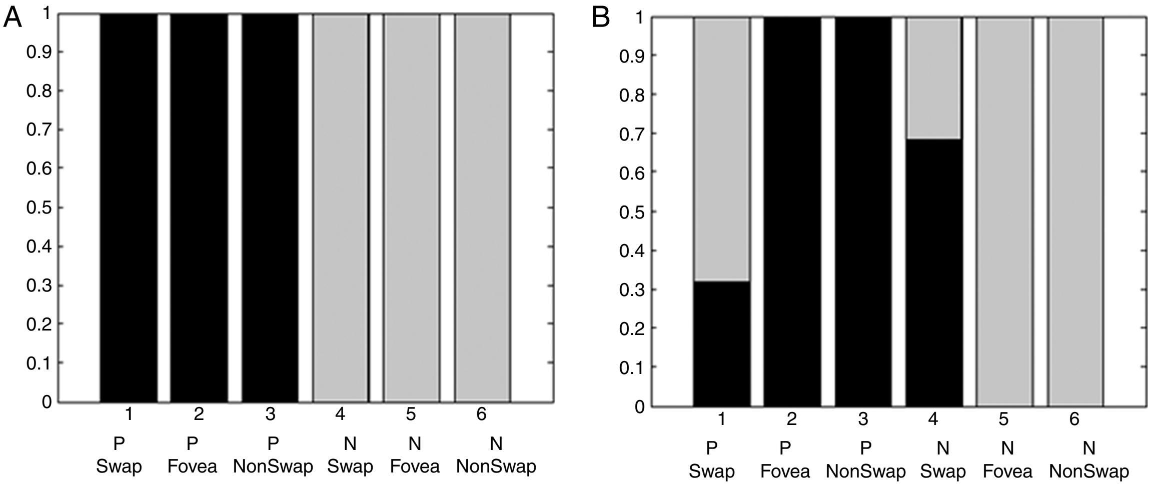


**Fig. 6.** Learned connection weights of six view category integrator neurons. Blacks denote weight strengths associated to object category *P*, and grays denote weight strengths associated to object category *N*. (A) Initially, object *P*'s views on all three retinal positions (Swap, Fovea, and NonSwap) are associated with object category *P*, and object *N*'s views on all three retinal positions are associated with object category *N*. (B) After 1000 swaps, the weights for both objects at swap positions (*P* Swap and *N* Swap) are greatly altered, while the weights at foveal and non-swap positions remain unchanged.

and simulated IT neurons. In comparison, 1000 normal exposures for the control non-swap position were also simulated. The selectivities of both neurons $O_p$ and $O_n$ at non-swap position show no visible change (Fig. 6).

In the Li and DiCarlo experiment, when the saccade brings an object's image to the fovea from an extra-foveal position, attention does not switch to other objects. The attentional shroud of an object remains active during an eye movement towards a stationary object because the shroud corresponding to the object is computed in head-centered coordinates. Thus, the shroud remains on during each individual normal or swap exposure. As a result, in the simulations, the shroud turned on when an exposure starts and off when an exposure ends.

## 4. Discussion

*Target swapping fools the reset mechanism*. The pARTSCAN model proposes how the brain can learn object categories that are invariant across object positions, sizes, and views. Key model mechanisms that enable position-invariant object learning enable the model to quantitatively simulate the Li and DiCarlo (2008) swapping data. In effect, the Li and DiCarlo (2008) experiments bypass the mechanism whereby attentional shrouds normally get reset when one object is replaced by another one. Their results illustrate a failure both of position-invariant *and* view-invariant category learning.

Attentional shrouds can also combine the views of an object as it is seen at different distances, thereby creating images of different sizes on the retina. An attentional shroud for the object will remain active during these continuous changes in the object's retinal image, and thereby enable views that vary in size to be combined through learning at the same invariant object category. Indeed, Li and DiCarlo (2010) have recently described a variant of the target swapping experiment in which targets of different sizes are swapped and the corresponding mixtures of objects are learned that would be expected from fooling the shroud mechanisms once again.

*fMRI test of reset mechanism during target swapping in humans*. Target swapping experiments can also be carried out in humans. If the swapping procedure does indeed fool the shroud reset mechanisms, and if the Chiu and Yantis (2009) fMRI data on transient parietal reset bursts reflects shroud and category reset, then no reset burst should occur in swapping tasks that can recode object categories. In addition, inserting large enough delays between the swapped objects should again cause such transient reset bursts to occur, as well as the learning of separable object categories. Thus, variations of the swapping paradigm can be used to dissociate the parietal regions that maintain sustained spatial attention on an object, using shrouds, versus those that shift attention to new objects, using transient reset bursts (Corbetta et al., 2000; Yantis et al., 2002). The predicted interactions between these two types of mechanism, and their interpretation in the swapping paradigm, can provide a new experimental probe in both monkeys and humans for studying the role of the parietal cortex in episodic memory and memory retrieval (Cabeza et al., 2008; Ciaramelli, Grady, & Moscovitch, 2008).

*Reversible disabling of reset mechanism using TMS*. It would be of interest if the shroud reset mechanism could be reversibly inactivated, say by transcranial magnetic stimulation, or TMS, leading to learned merging of view categories from more than one object into a single object category in cases that these view categories would otherwise be separated into distinct object categories.

*A learning mechanism for survival of unexpected dangers*. One might imagine that the most memory-saving method for recording position-invariant object information might be for the brain to somehow automatically transfer extra-foveal views into the fovea to learn invariant object categories. Then only foveal views would need to be learned. The Li and DiCarlo data shows that this is not true. Position-specific information has to be learned. Otherwise, there would be no difference between the swap and non-swap positions in the Li and Dicarlo data.

Why does the brain not use the potentially most memory-saving recoding method? It may be due to survival demands in a dangerous world. Transferring views takes time. Direct and rapid recognition of invariant object categories by extra-foveal views may save an animal's life in case a dangerous enemy is approaching. The model suggests that extra machinery is needed to learn position-invariant object categories and to learn stronger associations with more predictive unambiguous object views than less predictive ambiguous views, above and beyond the spatial attentional modulation that is needed to learn view-invariant object categories located at or near the fovea. View integrator neurons and competitive outstar learning are predicted to contribute to these additional competences. Indeed, in the Li and DiCarlo (2008) experiment, objects *P* and *N* become "ambiguous views" in the swap position (see Fig. 6(B)).

Various methods have been suggested for modeling neurons with persistent activities (Brunel, 2003; Durstewitz, Seamans, & Sejnowski, 2000; McCormick, 2001; Mongillo, Amit, & Brunel, 2003; Mongillo, Barak, & Tsodyks, 2008; Wang, 2001). We have used the simplest model, a leaky integrator with slow decay. Shunting dynamics with slow decay has the conceptual advantage that neuron activity has a fixed upper bound, but the result will not change for the simulation of the Li and DiCarlo data.

These results suggest that any object recognition model which does not have processes such as invariant object category reset by spatial attentional collapse, persistent activity in IT, and competitive outstar learning may have difficulty in explaining the Li and DiCarlo data, assuming that such a model carries out unsupervised incremental learning in real time, and does not impose biologically unrealistic hypotheses. For example, the HMAX model (Riesenhuber & Poggio, 1999; Serre, Kreiman et al., 2007; Serre, Oliva, & Poggio, 2007), which uses a MAX operation through several cortical regions to derive position invariance, has none of these features. It will be of interest to explore alternative possible explanations of swapping data, and their different predictions.

The current model is also being applied to image processing applications using naturally occurring objects. For example, preliminary results describe a neural model for solving the Where's Waldo problem (Chang, Cao, & Grossberg, 2009), or how the brain can efficiently search for, and learn to recognize, a desired target in a cluttered natural scene.

## 5. Methods

### 5.1. Model equations

*Retina/LGN: discounting the illuminant and contrast normalization*. The luminance of the retinal input image $I_{pq}$ at position $(p, q)$ is preprocessed by the model retina/LGN to discount the illuminant and contrast-normalize the image using shunting on-center off-surround networks (Grossberg & Todorovic, 1988). The equilibrium output signals $X_{ij}^{+}$ and $X_{ij}^{-}$ of ON and OFF cells, respectively, at position $(i, j)$ are defined by

$$X_{ij}^{+} = \left[X_{ij} - 0.05\right]^{+}, \tag{1}$$

$$X_{ij}^{-} = \left[-X_{ij} - 0.05\right]^{+}, \tag{2}$$

where

$$X_{ij} = \frac{4(C_{ij} - S_{ij})}{10^{-5} + C_{ij} + S_{ij}}, \tag{3}$$

notation $[w]^{+} = \max(w, 0)$ defines a threshold-linear output signal function, $C_{ij}$ is the Gaussian on-center input:

$$C_{ij} = \sum_{pq} I_{pq} G_{pqij}^{c}, \tag{4}$$

$S_{ij}$ is the Gaussian off-surround input:

$$S_{ij} = \sum_{pq} I_{pq} G_{pqij}^{s}, \tag{5}$$

and $G_{pqij}^{v}$ is a Gaussian kernel:

$$G_{pqij}^{v} = N^{v} \exp\left(-\frac{(p-i)^2 + (q-j)^2}{2\sigma_v^2}\right), \quad v = c, s. \tag{6}$$

Constant $N^{v}$ in (6) is chosen so that $N^{v} \sum_{pq} G_{pqij}^{v} = 1$. The width of the on-center and the off-surround are determined in (6) by $\sigma_c = 0.3$ and $\sigma_s = 2$, respectively.

*Cortical magnification using a log-polar transformation*. The ON and OFF cell output signals in (1) and (2) undergo a log-polar transformation that maps from retina position $(p, q)$ to cortex position $(x, y)$, defined by

$$M = 7\log(Z + 0.3), \tag{7}$$

where $M$ and $Z$ are complex numbers such that $M = x + iy$, and $Z = p + iq$. Eq. (7) transforms the retinal image into polar coordinates, and models the cortical magnification factor in humans and other primates (Schwartz, 1980). The outputs of this operation are log-polar maps $L^{+}$ and $L^{-}$.

*Object boundary*. Object boundaries are represented by complex cells in the primary visual cortex. The model computes the object boundary activity $B_{ij}$ at position $(i, j)$ by

$$B_{ij} = L_{ij}^{+} + L_{ij}^{-}, \tag{8}$$

where $L_{ij}^{+}$ and $L_{ij}^{-}$ are the log-polar transformed ON and OFF cell output signals.

*View category learning*. View category neurons receive input from object boundary neurons, and learn to respond to changes in position, size, and view. View categories are learned using mechanisms of Adaptive Resonance Theory, or ART (Carpenter & Grossberg, 1987, 1991; Grossberg, 1980). In particular, fuzzy ART (Carpenter, Grossberg, Markuzon, Reynolds, & Rosen, 1992; Carpenter, Grossberg, & Rosen, 1991) was chosen to simulate the learning of view category neurons because it simplifies the nonlinear dynamics of neural category learning into a more easily computable algorithm.

*Normalization*. The input boundary vector $B = \left(B_{ij}\right)$ is first preprocessed with complement coding that represents simultaneous processing by normalized ON and OFF cells:

$$E = \left(g(B_{ij}),\ 1 - g(B_{ij}) : \forall i, j\right), \tag{9}$$

where $g$ is a sigmoid signal function that normalizes all boundary signals to be less than 1:

$$g(x) = \frac{x^2}{0.25 + x^2}. \tag{10}$$

If $B$ is an $n$-dimensional vector of boundary positions, then $E$ is a "complement-coded" $2n$-dimensional vector of ON and OFF cell responses to the boundary.

*Category choice*. A neuron is called *committed* if it has learned to code one or more views. Otherwise, it is called *uncommitted*. The activity $V_j$ of the $j$th committed view category neuron in response to the complement-coded boundary input $E$ obeys:

$$V_j = \frac{\left|E \wedge W_j^{BV}\right|}{10^{-5} + \left|W_j^{BV}\right|}, \tag{11}$$

where $W_j^{BV}$ is the learned weight vector for the $j$th view category neuron, the fuzzy AND operator $\wedge$ is defined by $(x \wedge y)_i = \min(x_i, y_i)$, and the $L_1$ norm $|\bullet|$ is defined by $|x| = \sum_i |x_i|$. The minimum operation in the numerator may be interpreted as the expected number of learned sites that are activated by the input vector $E$. The most highly activated view category wins the competition among all active committed view neurons; that is, the $J$th committed neuron is chosen as the winner if

$$V_J = \max_j \{V_j : V_j > 0\}. \tag{12}$$

*Resonance or reset*. *Resonance* occurs if the chosen view category meets the matching criterion:

$$\frac{\left|E \wedge W_J^{BV}\right|}{|E|} \geq \rho, \tag{13}$$

where $\rho$ is the *vigilance* parameter that determines the network sensitivity to the match of bottom-up boundary $E$ and the learned top-down expectation with weights $W_J^{BV}$. Inequality (13) is computed in the orienting system (Fig. 3). It means that the amount of inhibition $\left|E \wedge W_J^{BV}\right|$ from the matched feature-expectation pattern exceeds the total excitation $|E|$ due to the input pattern $E$, multiplied by the vigilance, which plays the role of the gain of the excitatory input pattern $E$. In the current simulations, vigilance is chosen high at the value $\rho = 0.99$ because the Li and DiCarlo experimental procedure actively engages a monkey's attention.

When (13) occurs, the orienting system is inhibited and enables resonance to occur between the active category and the attended critical feature pattern, thereby triggering both category learning and learning of the top-down expectation.

If (13) is not satisfied, *mismatch reset* occurs. This happens because excitation from the bottom-up input $E$ exceeds inhibition from the matched feature-expectation pattern. Then a novelty-sensitive nonspecific arousal burst occurs from the orienting system and resets the currently active category (Fig. 3). As a result, the previously active view neuron $J$ is reset to inactive and a new winner is chosen by Eq. (12). The search process continues until the chosen winner satisfies (13).

*Learning*. If resonance occurs, then the winning category neuron $J$ learns to update its weight vector by the equation

$$W_J^{BV(\text{new})} = \beta(E \wedge W_J^{BV(\text{old})}) + (1-\beta)W_J^{BV(\text{old})}, \tag{14}$$

where the learning rate $\beta$ is set to 1 for fast learning. Otherwise, if no committed view category is active, an uncommitted view neuron is chosen as the winner $J$, thereby becomes committed, and learns to update its weight vector by the equation

$$W_J^{BV(\text{new})} = E. \tag{15}$$

*Output signals*. The output signals from view category neurons to view category integrator neurons are

$$V_j = \begin{cases} \dfrac{\left|E \wedge W_J^{BV(\text{new})}\right|}{10^{-5} + \left|W_J^{BV(\text{new})}\right|}, & \text{if } j = J, \\ 0, & \text{if } j \neq J. \end{cases} \tag{16}$$

*View category integrator*. As described in Section 2, view category integrator neurons preserve the activities of view category neurons. View category neurons may shut off when new views appear due to either object motion or eye movements, but their corresponding integrator neurons preserve their activities as long as the attentional shroud is on. The activity $U_i$ of the integrator neuron associated with view category neuron $i$ is defined by

$$\frac{\mathrm{d}U_i}{\mathrm{d}t} = -\alpha U_i + [V_i]^+ - U_i R, \tag{17}$$

where $\alpha$ is a small positive number ($\alpha \approx 0$) representing a slow decay rate, $V_i$ is the activity of view category neuron $i$, and $R$ is a reset signal controlled by a spatial attentional shroud (Fazl et al., 2009), here simply defined by

$$R = \begin{cases} 0, & \text{when shroud is on,} \\ R^*, & \text{when shroud is off,} \end{cases} \tag{18}$$

where $R^*$ is chosen large enough to rapidly inhibit the view category integrator neuron. As noted in Section 2, the shroud turns on when an exposure starts and off when an exposure ends. It does not shut off during a swap. Approximating $\alpha$ with 0 during each exposure, (17) can be solved as

$$U_i^{\text{new}} = U_i^{\text{old}} + \tau\,[V_i]^+, \tag{19}$$

where $\tau$ is the duration that an object image stays on at a particular retinal position; in other words, the duration that view $i$ is active between eye movements. In the simulations, $\tau = 1$ when an object's image is at the fovea, and $\tau = 0.5$ when it is at an extra-foveal position. This is consistent with data in the Li and DiCarlo experiment showing that the monkey foveated an object approximately twice long as it stays in extra-foveal positions (about 200 ms versus 100 ms; see the Supplementary Material of Li and DiCarlo (2008), page 3).

*Spatially-invariant object category learning*. Each invariant object category neuron with activity $\overset{\wedge}{O_j}$ receives summed inputs from all active view category integrator neurons:

$$\hat{O}_j = \sum_i \lambda_i W_{ij}^{UO}\,[U_i]^+, \tag{20}$$

where $\lambda_i$ is a bias parameter to the fovea with $\lambda_i > 1$ (using 10 in the simulation) when $U_i$ is an input from a foveal image and $\lambda_i = 1$ when $U_i$ is an input from an extra-foveal image, $W_{ij}^{UO}$ is the learned weight between view category integrator neuron $i$ and object category neuron $j$, and $U_i$ is the activity of view category integrator neuron $i$. Object category neurons compete with each other through a recurrent shunting on-center off-surround network, which normalizes the activity of each neuron $j$ (Grossberg, 1973). This rapid normalization process is approximated by

$$O_j = \frac{\hat{O}_j}{\sum_k \hat{O}_k}. \tag{21}$$

During learning, all weights $W_{ij}^{UO}$ are updated through an outstar learning equation (Grossberg, 1968, 1980):

$$\frac{\mathrm{d}W_{ij}^{UO}}{\mathrm{d}t} = a\,[U_i]^+ \left(f(O_j) - W_{ij}^{UO}\right), \tag{22}$$

where $a$ is a learning rate parameter, and $f$ is a sigmoid signal function defined by

$$f(O_j) = \begin{cases} 0, & \text{if } O_j \leq 0.1, \\ O_j, & \text{if } 0.1 < O_j < 0.9, \\ 1, & \text{if } O_j \geq 0.9. \end{cases} \tag{23}$$

When a small time step is used, (22) can be written in its discrete form as

$$W_{ij}^{UO(\text{new})} = W_{ij}^{UO(\text{old})} + \alpha\,[U_i]^+ \left(f(O_j) - W_{ij}^{UO(\text{old})}\right), \tag{24}$$

where $\alpha$ is the learning rate with $0 < \alpha \leq 1$. In the simulation, $\alpha = 0.0046$ best fits with the experimental data: the object category neuron ($O_j$) response to object images $P$ and $N$ at swap position is reversed at about 600 swaps (see Fig. 1). Increasing the learning rate will advance the reversal time, and vice versa (see Fig. 7).

*Simulation procedure and algorithm*. Based on the above model equations, the following algorithm was used to simulate the Li and DiCarlo (2008) data. The same procedure also provides an algorithm to learn spatially-invariant object recognition categories.

A. Initialization.

Set all weights for view category and object category neurons to 0.

B. For each normal or swap exposure do steps 1–3:

1. Set all view category integrator neurons and object category neurons to activity 0.
2. For each retinal image (where the object is in an extra-foveal position or the fovea), do steps 2.1–2.5:
   2.1. Compute the contrast enhancement signals, by Eqs. (1)–(6).
   2.2. Do log-polar transformations, by Eq. (7).

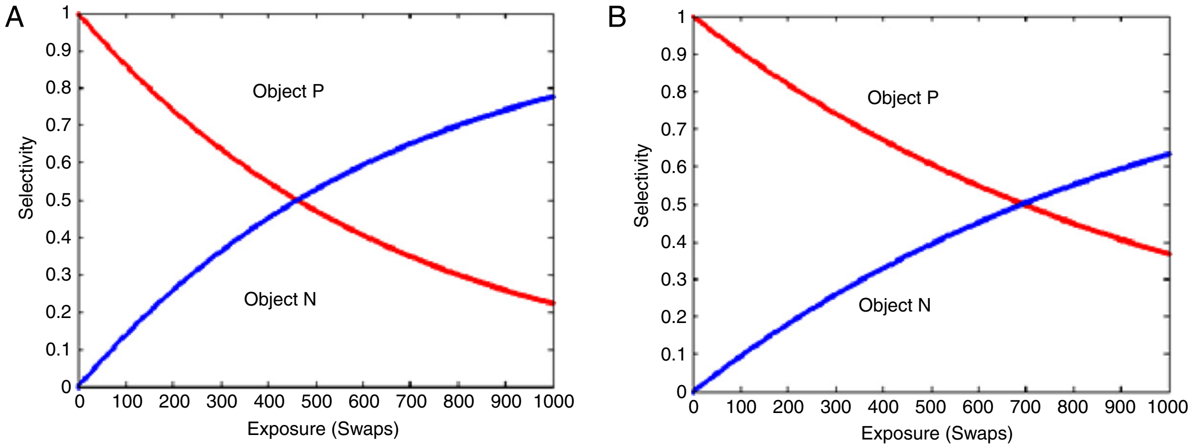


**Fig. 7.** Model simulations when the learning rate $\alpha$ in (24) varies. (A) Increasing the learning rate to $\alpha = 0.006$ advances the selectivity reversal time. The selectivity of model IT object category neuron $O_p$ to views of objects $P$ and $N$ at swap position reverses at about 450 swaps. (B) Decreasing the learning rate to $\alpha = 0.004$ delays the selectivity reversal time. The selectivity reverses at about 700 swaps now.

2.3. Compute the object boundary map, by Eq. (8).
2.4. Compute the view category neuron activities and update their weights using the fuzzy ART classifier, by Eqs. (9)–(16).
2.5. Compute the view category integrator neuron activities, by Eq. (19).

3. Compute the object category activities and update their weights, by Eqs. (20), (21), (23) and (24). In particular, in Eq. (20), if $\hat{O}_j = 0$ for all learned object category neurons $j$, a new object category neuron $J$ is activated. Its weight is then updated according to (24) with $f(O_J) = 1$.

## Acknowledgments

This work was supported in part by CELEST, a National Science Foundation Science of Learning Center (SBE-0354379), and by the SyNAPSE program of DARPA (HR0011-09-C-0001).

## Appendix

**Ambiguous View Learning Theorem.** *Consider Eqs. (20)–(22), where signal function $f$ is nonnegative with $f(0) = 0$, $f(1) = 1$, and $j = 1, \ldots, N$ are all object categories to which view $i$ is associated.*

(i) *If $f$ satisfies $\sum_j f(O_j) \leq f\left(\sum_j O_j\right)$, then $0 \leq \sum_j W_{ij}^{UO} \leq 1$ at equilibrium.*
(ii) *When $f$ is defined by (23), then $0 \leq \sum_j W_{ij}^{UO} \leq 1$ at equilibrium.*
(iii) *Assuming that $f$ is defined by (23) and $N = 2$, then $\sum_j W_{ij}^{UO} = 1$ at equilibrium, and furthermore $\sum_j W_{ij}^{UO} \equiv 1$ for all times $t$ if $\sum_j W_{ij}^{UO} = 1$ at time 0.*
(iv) *Assuming that $N = 1$, then $W_{i1}^{UO} = 1$ at equilibrium.*

**Proof.** (i) According to Eq. (22), we have

$$\frac{\mathrm{d}\sum_j W_{ij}^{UO}}{\mathrm{d}t} = a[U_i]^+\left(\sum_j f(O_j) - \sum_j W_{ij}^{UO}\right). \tag{A.1}$$

From Eq. (21), $\sum_j O_j = 1$. Therefore, $\sum_j f(O_j) \leq f\left(\sum_j O_j\right) = f(1) = 1$, which implies that, at equilibrium, $\sum_j W_{ij}^{UO} \leq 1$. It is obvious that $\sum_j W_{ij}^{UO} \geq 0$, since $f$ is nonnegative.

(ii) Case A: there is some $j$ such that $O_j \geq 0.9$. Then, for all other $j$, $O_j \leq 0.1$, since $\sum_j O_j = 1$ by (21). From (23), $\sum_j f(O_j) = 1$. This implies that (ii) holds.

Case B: For all $j$, $O_j < 0.9$. Then, by (23), $\sum_j f(O_j) \leq \sum_j O_j = 1$. This again implies that (ii) holds.

(iii) Case A: there is some $j$ such that $O_j \geq 0.9$. We again have $\sum_j f(O_j) = 1$, as the proof in (ii). Case B: For both $j = 1$ and 2, $O_j < 0.9$. This implies that $0.1 < O_j < 0.9$ for both $j$. Therefore, $\sum_j f(O_j) = \sum_j O_j = 1$ by (23). As a result, $\sum_j W_{ij}^{UO} = 1$ at equilibrium. It is then easy to see that $\sum_j W_{ij}^{UO} \equiv 1$ for all time $t$ if $\sum_j W_{ij}^{UO} = 1$ at time 0, since $\frac{\mathrm{d}\sum_j W_{ij}^{UO}}{\mathrm{d}t} = 0$.

(iv) Since $N = 1$, we have $O_1 = 1$ from (21). Thus $f(O_1) = 1$. This implies that $W_{i1}^{UO} = 1$ at equilibrium from (22). □

**Remarks.** The Ambiguous View Learning Theorem tells us that, when an object view is associated with more than one object category, the sum of all its learned weights cannot be greater than 1, which is the learned weight when a view is associated to only one object category. In other words, distinguishable views learn stronger associations with object category neurons than do ambiguous views.

## References

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

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

Booth, M. C., & Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex*, *8*, 510–523.

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

Brunel, N. (2003). Dynamics and plasticity of stimulus selective persistent activity in cortical network models. *Cerebral Cortex*, *13*, 1151–1161.

Bulthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(1), 60–64.

Bulthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex*, *5*(3), 247–260.

Cabeza, R., Ciaramelli, E., Olson, I. R., & Moscovitch, M. (2008). The parietal cortex and episodic memory: an attentional account. *Nature Reviews Neuroscience*, *9*, 613–625.

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.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern-recognition machine. *Computer Vision, Graphics, and Image Processing*, *37*, 54–115.

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*, 131–137.

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., & Rosen, D. B. (1992). Fuzzy ARTMAP—a neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*, *3*, 698–713.

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*, 759–771.

Carpenter, G. A., & Ross, W. D. (1993). ART-EMAP: a neural network architecture for learning and prediction by evidence accumulation. In *Proceedings of the world congress on neural networks. WCNN-93. III* (pp. 649–656).

Chang, H.-C., Cao, Y., & Grossberg, S. (2009). Where's Waldo? How the brain earns to categorize and discover desired objects in a cluttered scene [Abstract]. *Journal of Vision*, *9*(8), 173.

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*, 3930–3938.

Ciaramelli, E., Grady, C. L., & Moscovitch, M. (2008). Top-down and bottom-up attention to memory: a hypothesis (AtoM) on the role of the posterior parietal cortex in memory retrieval. *Neuropsychologia*, *46*, 1828–1851.

Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P., & Shulman, G. L. (2000). Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature Neuroscience*, *3*, 292–297.

Daniel, P. M., & Whitteridge, D. (1961). The representation of the visual field on the cerebral cortex in monkeys. *The Journal of Physiology*, *159*, 203–221.

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

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *353*(1373), 1245–1255.

Desimone, R., & Gross, C. G. (1979). Visual areas in the temporal cortex of the macaque. *Brain Research*, *178*, 363–380.

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature Neuroscience*, *3*, 1184–1191.

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.

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*, 1–48.

Fischer, H. (1973). Overlap of receptive field centers and representation of the visual field in the cat's optic tract. *Vision Research*, *13*, 2113–2120.

Foley, N. C., Grossberg, S., & Mingolla, E. (2011). Neural dynamics of object-based multifocal visual spatial attention and priming: object cueing, useful-field-of-view, and crowding (submitted for publication).

Fuster, J. M., & Jervey, J. P. (1981). Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. *Science*, *212*, 952–955.

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

Grossberg, S. (1968). Some nonlinear networks capable of learning a spatial pattern of arbitrary complexity. *Proceedings of the National Academy of Sciences*, *59*, 368–372.

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

Grossberg, S. (1976). Adaptive pattern classification and universal recoding, I: parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121–134.

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

Grossberg, S. (1987). Cortical dynamics of three-dimensional form, color, and brightness perception, II: binocular theory. *Perception & Psychophysics*, *41*, 117–158.

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*, *20*(9), 1040–1053.

Grossberg, S. (2009). Cortical and subcortical predictive dynamics and learning during perception, cognition, emotion, and action. *Philosophical Transactions of the Royal Society of London*, *364*, 1223–1234.

Grossberg, S., & Huang, T. (2009). ARTSCENE: a neural system for natural scene classification. *Journal of Vision*, *9*, 1–19.

Grossberg, S., & Mingolla, E. (1985). Neural dynamics of perceptual grouping: textures, boundaries, and emergent segmentations. *Perception & Psychophysics*, *38*(2), 141–171.

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

Grossberg, S., & Todorovic, D. (1988). Neural dynamics of 1-D and 2-D brightness perception: a unified model of classical and recent phenomena. *Perception & Psychophysics*, *43*, 241–277.

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

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

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*, 816–824.

Ito, M., Tamura, H., Fujita, I., & Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, *73*, 218–226.

Kohonen, T. (1989). *Self-organization and associative memory* (3rd ed.) Berlin: Springer-Verlag.

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), 406–413.

Li, N., & DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science*, *321*, 1502–1507.

Li, N., & DiCarlo, J. J. (2010). Unsupervised natural visual experience rapidly reshapes size invariant object represent in inferior temporal cortex. *Neuron*, *67*, 1062–1075.

Logothetis, N. K., Pauls, J., Bulthoff, H. H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology*, *4*(5), 401–414.

McCormick, D. A. (2001). Brain calculus: neural integration and persistent activity. *Nature Neuroscience*, *4*, 113–114.

Miller, E. K., Li, L., & Desimone, R. (1993). Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *Journal of Neuroscience*, *13*, 1460–1478.

Miyashita, Y., & Chang, H. S. (1988). Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature*, *331*, 68–70.

Mongillo, G., Amit, D. J., & Brunel, N. (2003). Retrospective and prospective persistent activity induced by Hebbian learning in a recurrent cortical network. *European Journal of Neuroscience*, *18*, 2011–2044.

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic theory of working memory. *Science*, *319*, 1543–1546.

Otto, T., & Eichenbaum, H. (1992). Neuronal activity in the hippocampus during non-match to sample performance in rats: evidence for hippocampal processing in recognition memory. *Hippocampus*, *2*, 323–334.

Peterhans, E., & von der Heydt, R. (1989). Mechanisms of contour perception in monkey visual cortex. II. Contours bridging gaps. *Journal of Neuroscience*, *9*, 1749–1763.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize 3D objects. *Nature*, *343*, 263–266.

Pollen, D. A. (1999). On the neural correlates of visual perception. *Cerebral Cortex*, *9*(1), 4–19.

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), 100–113.

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*, 1019–1025.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, *3*, 1199–1204.

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

Schwartz, E. L. (1980). Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding. *Vision Research*, *20*, 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 of the United States of America*, *80*, 5776–5778.

Seibert, M., & Waxman, A. M. (1992). Adaptive 3-D object recognition from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *14*(2), 107–124.

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), 6424–6429.

Spitzer, H., Desimone, R., & Moran, J. (1998). Increased attention enhances both behavioral and neuronal performance. *Science*, *240*, 338–340.

Tomita, H., Ohbayashi, M., Nakahara, K., Hasegawa, I., & Miyashita, Y. (1999). Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature*, *401*, 699–703.

Tootell, R. B., Silverman, M. S., Switkes, E., & DeValois, R. L. (1982). Deoxyglucose analysis of retinotopic organization in primate smote cortex. *Science*, *218*, 902–904.

Tyler, C. W., & Kontsevich, L. L. (1995). Mechanisms of stereoscopic processing: stereoattention and surface perception in depth reconstruction. *Perception*, *24*(2), 127–153.

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.

Vinogradova, O. S. (1975). Functional organization of the limbic system in the process of registration of information. Facts and hypotheses. In R. L. Isaccson, & K. H. Pribram (Eds.), *The hippocampus*: *Vol. 2* (pp. 3–69). Plenum Press.

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.

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

Wang, X.-J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neurosciences*, *24*, 455–463.

Werblin, F. S. (1971). Adaptation in a vertebrate retina: intracellular recordings in Necturus. *Journal of Neurophysiology*, *34*, 228–241.

Yantis, S., Schwarzbach, J., Serences, J. T., Carlson, R. L., Steinmetz, M. A., Pekar, J. J., et al. (2002). Transient neural activity in human parietal cortex during spatial attention shifts. *Nature Neuroscience*, *5*, 995–1002.

Yarbus, A. F. (1967). *Eye movements and vision*. New York: Plenum Press.