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.