Commentary

Visual object recognition: building invariant representations over time

We easily and rapidly recognize visual objects, for example a chair, from a variety of viewpoints, regardless of their position on the retina and under a great diversity of image conditions (Thorpe *et al* 1996). This extraordinary flexibility of object perception is a key unresolved question in vision research and computational neuroscience. Our lack of understanding of how object recognition works is highlighted by the poor success of computer vision systems in approaching levels of performance exhibited by biological vision (Pinto *et al* 2008). In spite of decades of research, this fact remains true even for arguably well-constrained tasks, such as inspecting luggage at airports. The challenge is that any object can produce, essentially, an infinite number of retinal images, requiring a recognition system whose final output is unaffected by this vast input variation. Consequently, object recognition requires a system that is functionally invariant to changes in scale, viewing angle, retinal position, luminance, contrast, motion and other image transformations. (The real problem is actually even more complicated, as the "same" objects often come in many different shapes.)

While proving difficult for computer vision, this problem is easily solved by neurons in the inferior temporal (IT) cortex — a high level visual area that generally exhibits object selective responses that are tolerant to changes in position, scale and viewing angle (Ito *et al* 1995; Logothetis and Pauls 1995). However, little is known about how such invariances are formed. Addressing this significant question, two recent neurophysiological studies by James DiCarlo and colleagues focused on position invariance — the ability to recognize objects irrespective of retinal position (Kravitz *et al* 2008). Development of positional invariance is needed as objects projected at different retinal locations stimulate different populations of neurons and might be represented at very different levels of detail, depending on the distance from high-resolution central vision. One possibility is that position invariance is automatic: what is learned about an object at one location automatically generalizes to the entire visual field (Olshausen *et al* 1993). Alternatively, position invariance could build up slowly as a particular object is seen at different retinal locations. The latter strategy would benefit from saccadic eye movements, which cause positional shifts of the retinal image several times per second. Thus, with each saccade, "temporal contingency" leads to the same object being placed at a different location.

In a recent study published in *Science*, Li and DiCarlo (2008) investigated whether temporal contiguity of object experience across saccades shapes position invariance of neurons in monkey IT cortex. This hypothesis was tested by creating an altered world in which temporal contiguities were modified. For each neuron, the authors first identified two objects, one yielding a strong response, such as a car, and another that was less preferred, e.g. a chair. On each trial, the preferred object was first presented in one of two peripheral locations. Typically, that caused a monkey to quickly saccade to the visible object — a natural exploratory primate behavior. At one of the two locations, the display remained constant as the monkey looked toward the object. However, at the other location, the presented object was replaced (swap exposure) with the less preferred object as the monkey moved its eyes (this ensured that the swap was inconspicuous because the visual system is essentially blind while saccades are executed) (figure 1). Following several 15 min training sessions in this artificial visual world, the authors reexamined neural selectivity for the two initially selected objects. The results showed that IT neurons retained their initial object preference at all locations except that where the "swap" occurred. At this location, the neurons exhibited an increased response to the initially less preferred object and reduced response to

Keywords. Interferotemporal cortex; object invariance; object recognition; positional tolerance; saccadic eye movements

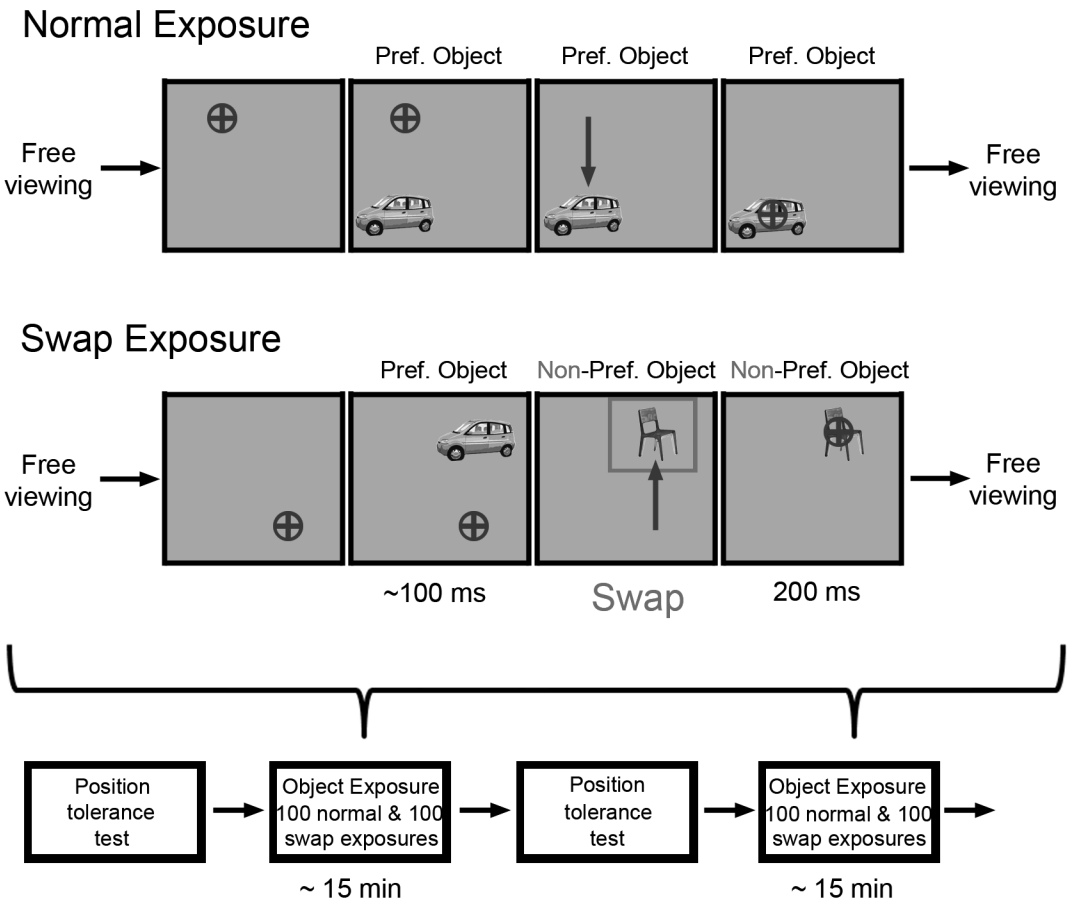

Figure 1. Schematic representation of the experimental design used by Li and DiCarlo. In their test, monkeys explored an altered visual world (free viewing, which consisted of viewing a video monitor where isolated objects occurred occasionally). Every ~15 min exploration was interrupted and IT neurons were tested for potential changes in position tolerance. This test was alternated with an object exposure phase that consisted of 100 "normal" (50 Preferred→50 Preferred, 50 Non-Preferred→50 Non-Preferred) and 100 "swap" exposures (50 Preferred→50 Non-Preferred, 50 Non-Preferred→50 Preferred). Position tolerance of object selectivity was tested through presentation of each object at 3° below, above and at the center of gaze. The crosses on the upper and lower panels indicate the center of gaze. Modified from Li and DiCarlo (2008).

the preferred object. This remarkable result shows that position invariance of IT neurons can be broken by a relatively brief exposure to an artificial world where the temporal contiguity of visual experience is systematically altered. Importantly, these changes of object preferences at the swap location occurred during unsupervised behavior, suggesting that this plasticity might reflect a natural process that occurs as we interact with objects. Does this result imply that the monkeys confused the "swapped" object for the initially preferred one? While the present study does not answer this question, previous behavioral experiments with human subjects have shown that exposure to similarly altered worlds can create object confusions. In these studies, object confusions were created by changing object identity in synchrony with viewpoint changes (Wallis and Bülthoff, 2001) or saccade-dependent changes in retinal position (Cox *et al* 2005). Moreover, by showing that no confusions occurred if subjects did not move their eyes, Cox *et al* (2005) demonstrated that the pairing of object changes with saccadic eye movements is necessary to generate position-dependent object confusions. The object changes in these human studies, however, were arguably less dramatic than those in Li and DiCarlo (2008).

By showing that the positional tolerance of IT neurons can be broken by altering temporal contiguity of visual experience, Li and DiCarlo's work suggests that it is our experience with objects at many

different locations that generates position invariance. This assertion predicts that learning to discriminate objects at one location will not automatically transfer to other visual field positions. This hypothesis too was tested in a recent study by Cox and DiCarlo, published in *The Journal of Neuroscience* (Cox and DiCarlo 2008). Here, the authors trained monkeys to discriminate four different shapes at a particular location in the visual field. After training, IT neurons showed increased selectivity for discriminated objects at the testing location. Interestingly, this selectivity did not transfer to an untrained location, indicating that positional invariance is not automatic; it requires exposure at various locations. The asymmetry in neural selectivity was mirrored in behavioral performance, which was significantly better at the trained location.

In these two studies, DiCarlo and colleagues have shown that IT neurons do not exhibit position invariance for novel objects automatically, and that existing positional invariance can be broken by altering temporal contiguity of visual object experience. The latter finding points to a pragmatic mechanism by which positionally invariant representation might develop in the IT cortex. As we visually scan our environment, each object's position in the visual field rapidly changes with each eye movement. This quickly builds up a succession of temporally contiguous exposures. Given that most objects are stationary and do not change appearance abruptly (as they do in the Li and DiCarlo 2008 study), visual input contains statistically useful information that can be used to build up positional tolerance. In fact, studies have repeatedly shown that primates are exceptionally good at exploiting such statistical regularities in a wide variety of tasks (e.g. Saffran *et al* 1996; Chun 2000).

An important question for future research is whether temporal contiguity plays a role in building up other invariant object representations, such as spatial scale and viewpoint. Or is positional tolerance a special case? Throughout the visual system, object features are typically represented in the retinotopic (i.e., eye-centered) coordinates (Chklovskii and Koulakov 2004). This physiological constraint might provide a bias towards positional invariance that relies on multiple spatially specific representations. Even large receptive fields in extrastriate visual areas can have independent sub-fields that are more specific with regard to spatial locations in the visual field (e.g. Livingstone *et al* 2001). This natural coupling of object position and visual system architecture does not easily extend to scale and viewpoint changes. On the other hand, active interaction with our environment provides abundant experience with temporally contiguous changes in object size and viewpoint. This alone might be sufficient for the development of object tolerances analogous to that which DiCarlo and colleagues describe for object position. Support for this hypothesis comes from Wallis and Bülthoff (2001), who showed that smooth temporally contiguous changes in object view are important in the development of invariance with respect to the viewing point.

In summary, these papers by DiCarlo and colleagues show that the brain might rely on sequences of temporally contiguous experiences with visual objects to gradually build up positionally tolerant representations. A similar mechanism might be employed in the development of other object invariances, such as those related to viewpoint and scale changes. Such a strategy might seem inefficient, especially when compared to a theoretical alternative where invariant representation is extrapolated from restricted experiences with objects (Olshausen *et al* 1993). It is possible, however, that such automatic generalization algorithms are simply too prone to errors and are not trivial to implement. This might explain why artificial vision systems pale in comparison to the robustness of biological vision. Alternatively, it can be argued that the input-driven computational strategy described by DiCarlo and colleagues is particularly efficient because it relies on the extraordinary amount of temporally contiguous object experiences we naturally receive. Saccadic movements alone provide 100 million sequential and orderly transformations of the visual input per year. Evolution would have been foolish to ignore this treasure chest of information.

References

Chklovskii D B and Koulakov A A 2004 Maps in the brain: what can we learn from them?; *Annu. Rev. Neurosci.* **27** 369–392

Cox D D, Meier P, Oertelt N and DiCarlo J J 2005 'Breaking' position invariant object recognition; *Nat. Neurosci.* **8** 1145–1147

Cox D D and Dicarlo J J 2008 Does Learned Shape Selectivity in Inferior Temporal Cortex Automatically Generalize Across Retinal Position?; *J. Neurosci.* **28** 10045–10055

Ito M, Tamura H, Fujita I and Tanaka K 1995 Size and position invariance of neuronal responses in monkey inferotemporal cortex; *J. Neurophysiol.* **73** 218–226

Kravitz D J, Vinson L D and Baker C I 2008 How position dependent is visual object recognition?; *Trends Cogn. Sci.* **12** 114–122

Li N and DiCarlo J J 2008 Unsupervised natural experience rapidly alters invariant object representation in visual cortex; *Science* **321** 1502–1507

Livingstone M S, Pack C C and Born R T 2001 Two-dimensional substructure of MT receptive fields; *Neuron* **30** 781–793

Logothetis N K and Pauls J P 1995 Psychophysical and physiological evidence for viewer-centered object representation in the primate; *Cereb. Cortex* **5** 270–288

Olshausen B A, Anderson C H and Van Essen D C 1993 A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information; *J. Neurosci.* **13** 4700–4719

Pinto N, Cox D D and DiCarlo J J 2008 Why is Real-World Visual Object Recognition Hard?; *PLoS Comput. Biol.* **4** e27

Thorpe S, Fize D and Marlot C 1996 Speed of processing in the human visual system; *Nature (London)* **381** 520–522

Wallis G and Bülthoff H H 2001 Effects of temporal association on recognition memory; *Proc. Natl. Acad. Sci. USA* **98** 4800–4804

Duje Tadin* and Raphael Pinaud**
Department of Brain and Cognitive Sciences and
Center for Visual Science,
University of Rochester,
Rochester, NY 14627, USA
**(Email, duje@cvs.rochester.edu); **(Email, pinaud@bcs.rochester.edu)*

*e*Publication: 13 November 2008