Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Current Opinion in Neurobiology

# Do we understand high-level vision?

David Daniel Cox

‘High-level’ vision lacks a single, agreed upon definition, but it might usefully be defined as those stages of visual processing that transition from analyzing local *image structure* to analyzing *structure of the external world* that produced those images. Much work in the last several decades has focused on object recognition as a framing problem for the study of high-level visual cortex, and much progress has been made in this direction. This approach presumes that the operational goal of the visual system is to read-out the identity of an object (or objects) in a scene, in spite of variation in the position [1], size [2], lighting [3] and the presence of other nearby objects [4,5]. However, while object recognition as a operational framing of high-level is intuitive appealing, it is by no means the only task that visual cortex might do, and the study of object recognition is beset by challenges in building stimulus sets that adequately sample the infinite space of possible stimuli. Here I review the successes and limitations of this work, and ask whether we should reframe our approaches to understanding high-level vision.

**Addresses**
Department of Molecular and Cellular Biology, Center for Brain Science, School of Engineering and Applied Sciences, Harvard University 52 Oxford St., Room 219.40, Cambridge, MA 02138, United States

Corresponding author: Cox, David Daniel (davidcox@fas.harvard.edu)

**Current Opinion in Neurobiology** 2014, **25**:187–193

This review comes from a themed issue on **Theoretical and computational neuroscience**

Edited by **Adrienne Fairhall** and **Haim Sompolinsky**

For a complete overview see the Issue and the Editorial

Available online 16th February 2014

0959-4388/$ – see front matter, Published by Elsevier Ltd.

**http://dx.doi.org/10.1016/j.conb.2014.01.016**

## A failure of intuition: why is vision so deceptively difficult?

We parse and understand our visual environments so effortlessly that is easy to overlook what an impressive computational feat this represents. The central computational challenge of vision arises from the fact that it is an ‘ill-posed’ problem — the external world is three dimensional and made up of surfaces with complex reflectance properties, yet the visual system must make do with a pair of two-dimensional retinas containing only a handful of photoreceptor types. Any given object can cast an effectively infinite number of different images onto the retina, depending on the object’s position relative to the viewer, the configuration of light sources, and the presence of other objects (Figure 1a). Conversely, any given image falling on the retina can theoretically correspond to infinitely many possible configurations of surfaces in the 3D world (Figure 1b shows an illusion that relies on this fact).

A ‘meta-problem’ also arises when we set out to *study* the visual system: our intuitions about vision are frequently wrong, because we perceive visual tasks to be easy when in fact they are quite hard. These tasks only seem easy because our visual system has evolved over millions of years to solve these complex tasks efficiently and effectively. In a particularly illustrative example, in the early days of artificial intelligence research, Seymour Papert famously proposed that an undergraduate student could ‘solve’ vision as a summer project.

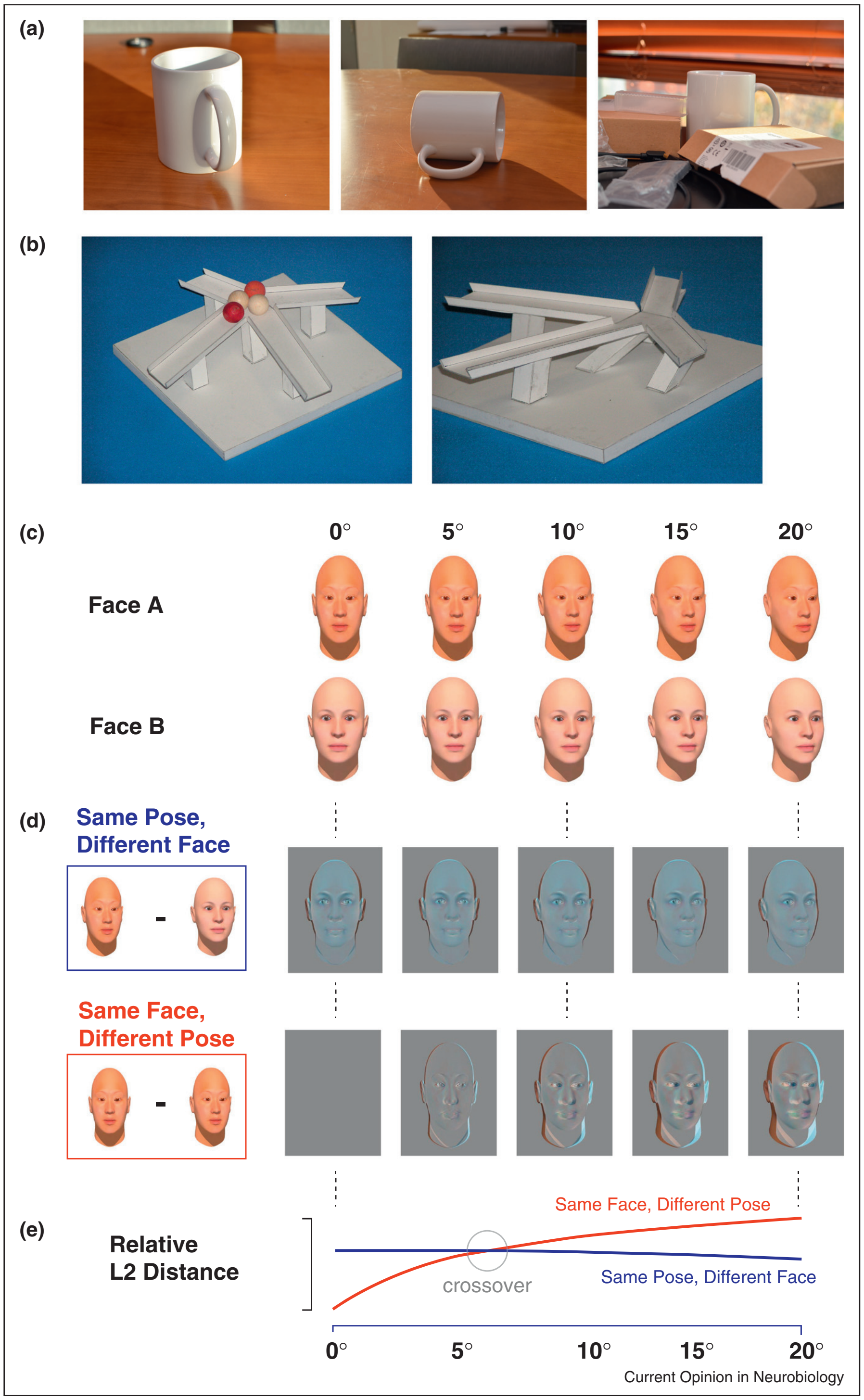
The competence of our visual systems also distorts our fundamental intuitions about the nature of the problem. Figure 1c illustrates the difference between perceptual space (i.e. how our brains perceive images) and pixel space: the same face, seen from slightly different angles, looks almost identical, whereas different individuals viewed from the same angle look obviously different. In a pixel-wise comparison of these image pairs (Figure 1d), different individuals viewed from the same angle are more similar than the same individual viewed from different angles, even for relatively small angles (Figure 1e). However, our visual systems allow us to see through these large differences in the raw two dimensional images on our retinas, to find the underlying commonality in the three dimensional objects that produced them.

## In search of an operational definition of ‘high-level’ vision

Because our intuitions in high-level vision can fail us in these ways, when we study high-level vision, it is critical to work from a concrete operational definition of the phenomenon we are studying. The theoretical issues raised above are often collectively grouped under the umbrella of ‘high-level’ vision, and while the term is frequently used, no single agreed-upon definition exists.

Anatomically, visual cortex in primates is known to be organized into two parallel pathways: a ‘ventral’ pathway thought to be involved primarily in processing visual form, and a ‘dorsal’ pathway thought to be more involved in analyzing visual motion and visual space. The ventral pathway is generally conceptualized as a hierarchical cascade, with visual inputs arriving from the thalamus in area V1, and then being successively processed in a cascade of areas (e.g. V2 to V4 to IT). A loose shorthand for this hierarchy is to refer to early stages of the cascade as ‘low-level’, and later stages as ‘high-level vision’. Many

**Figure 1**


(a)
(b)
(c)
0°
5°
10°
15°
20°
Face A
Face B
(d)
Same Pose, Different Face
-
Same Face, Different Pose
-
(e)
Relative L2 Distance
Same Face, Different Pose
crossover
Same Pose, Different Face
0°
5°
10°
15°
20°
Current Opinion in Neurobiology

Vision is an ill-posed problem. **(a)** Any given object in the world can cast infinitely many different images onto the retina, depending on its position relative to the viewer, lighting, and the presence of other objects. **(b)** Correspondingly, any given retinal image can theoretically correspond to infinitely many different objects in the real world. Here, an optical illusion that relies on this effect is shown (courtesy of Kokichi Sugihara [36]). While the left image looks like four upward-going ramps meeting in the center, this particular image corresponds to one of the many other possible real-world objects that can cast approximately the same retinal image. The right image shows the same object from a different view, revealing that the ramps actually all point downward. **(c)** Pixel-level image variation caused by variation in viewing parameters for single object is often larger than the pixel differences between different objects. Here we show 3D-rendered images of the faces of two individuals undergoing a rotation through 20° in azimuth. **(d)** If we look at pixel-wise difference images between the faces in the same pose but across individuals (blue) and between the same individual but in different poses (red), we see that the pixel differences induced by differences in pose outstrip the pixel differences caused by differences in identity. **(e)** Here we show the L2 (Euclidean) distance between each pair of images, plotted as a function of rotation angle. Pixel differences caused by pose variations of 5–10° are already greater than the pixel differences between individuals.

visual response properties conspicuously vary along the hierarchy of visual areas. As we move from V1 to IT, mean neuronal response latency increases, implying some measure of causality. Receptive field sizes increase dramatically, implying pooling and convergence of inputs from lower to higher-level areas. 'Complexity' of tuning also increases, implying integration of elaborate form tuning from simpler tuning found in lower levels.

A functional definition of 'high-level' vision focuses on the extent to which visual processing estimates physical properties of actual objects and surfaces in the environment, rather than being primarily concerned with measuring physical properties of light cast onto the retina. Physical properties of an object might include object identity (e.g. 'cup' or 'face'), properties of shape (spherical, pointed), and material properties ('metallic', 'soft'). Roughly speaking, high-level vision processing moves from describing the structure of the image, to describing the structure of the external world that produced that image — the surfaces, materials and light sources that *produced* the image, and their arrangement relative to one another. Like the anatomical definition, this functional definition forms a continuum rather than a clear dichotomy between 'low-level' and 'high-level'. Unlike the anatomical definition, this functional definition has the advantage of connecting to the goals of the visual system, allowing us to construct tests — both for natural and artifical (e.g. computer vision) systems — to explore that functionality.

For the last several decades, the task of interest in the study of high-level vision has been object recognition. This approach presumes the operational goal of the visual system is to read-out the identity of an object (or objects) in a scene, in spite of variation in the position [1], size [2], lighting [3] and the presence of other nearby objects [4,5].

## 'Reading-out' object identity

Over the past two decades, a general strategy of decoding or 'reading-out' population representations in visual cortical areas has emerged as an important tool for understanding higher-level visual cortex with both neuronal recordings in monkeys [6,7,8•] and in fMRI measurements of human object representations [9–12]. This approach implicitly recognizes that the problem of vision is not one of information content, but of format. We know that the activity of retinal ganglion cells contains all of the information that the visual system can act upon, and that nonlinearity and noise in neuronal processing can only decrease (and never increase) the absolute amount of information present. However, the information present in the firing of retinal ganglion cells is not in a format that can be easily read-out by a downstream neuron in order to guide action.

Linear classifiers are a reasonable choice of decoder to probe high-level neuronal codes (e.g. in areas V4, IT, and PRh) for extracting object identity information. A linear classifier is nothing more than a dot-product and a threshold, and while there is much debate about the fundamental computational abilities of single neurons, it is not hard to imagine a downstream neuron performing a weighted sum (integration of synaptically weighted inputs), followed by a threshold (the firing threshold of the downstream neuron). The strength of a linear readout approach — or any approach that pre-specifies a neurally plausible readout mechanisms — is that it allows us to test not just what information might be present, but what information could be theoretically read-out by a single downstream neuron. For instance, it has been shown that a simple downstream linear-plus-threshold neuron can theoretically decode object identity from a population of IT neurons, irrespective of object position and size [7].

Meanwhile, the field of computer vision has long used a similar operational definition of object recognition. While the goals of computer vision research are slightly different, given its focus on creating computational systems that have demonstrable performance in specific real-world applications, computer vision and visual neuroscience rely on a similar operational definition of the 'high-level vision' problem: within a given image, determine whether a given object is present or not. Over the past few years, computational models based on the ventral visual pathway have been developed; these models perform well against standard computer vision tasks, although they still fall far short of the performance of real biological visual systems.

## Challenges in exploring object recognition

While there is power in asking whether a neuronal population representation can support an object recognition task, there are three broad areas of difficulties inherent in this approach. First, defining objects is highly problematic. We could mean a particular object ('this specific chair'), or a class of object ('all chairs'). There are usually a large number of possible exemplars representing a given class, and individual exemplars can appear infinitely many ways due to variations in lighting and viewing angle. Some objects, like faces, have generally similar features (e.g. eyes, nose, mouth), but with very subtle but quite meaningful distinctions between them (e.g. close-set eyes, button nose). Other objects, like cups, may have widely different features (e.g. with or without handles). Still other objects, like chairs, may encompass a huge range of visually dissimilar objects (e.g. 'desk chair', 'folding chair', 'recliner'), and in fact be more effectively defined by their function (e.g. 'something you can sit on') rather than their visual appearance. We run a significant conceptual risk if we treat these aspects as if they were the same for all objects — but, we lack a principled way to draw the line between categories where semantic and visual content are aligned, versus those where they are dissociated.

Second, building stimulus sets that are representative of an object is deceptively difficult. As described earlier, any given 3D object can, depending on lighting, viewing angle, etc., cast an infinite array of 2D images on the retina — representing an infinite space of images to explore. Yet experimental limitations dictate that the number of stimuli we can show while characterizing a neuron is typically quite small. Emerging advanced recording technologies can help, but they cannot fully solve this problem — an organism in its full lifespan can only experience a comparatively small sample of the 'natural' images that are possible. Thus, when we set out to design an experiment, we must have not just an idea for what visual distinctions a given region of cortex might be good at representing, but we must also choose a finite set of stimuli that encapsulate this idea. Adequately sampling the space of possible stimuli presents a major difficulty in experimental design. Unfortunately, neurophysiology stimulus sets are almost always one-off, *ad hoc* sets, collected without a great deal of attention to these issues.

This problem is best illustrated by work done in the computer vision community on the Caltech 101 dataset [13], which served for many years as a benchmark for object recognition performance (e.g. [14–16]). It consists of images belonging to 101 object categories, and systems must categorize which object category appears in each picture. We [17] found that a very simple V1-inspired computational model was able to outperform all previous models on the Caltech 101 dataset, yet perform poorly on a conceptually simpler two-class problem which reflected more variation in lighting and viewing angle. This suggests that the standard test sets of representative images used to measure performance do not constitute a representative sample of those objects under real-world conditions, because they are typically curated images which are more posed, staged, and centered than would occur under real-world conditions. Similarly, Torralba and Efros [18••] studied multiple benchmark computer vision data sets that contained shared object categories (e.g. 'car'). They found that systems trained on one dataset generalized poorly to other datasets for the same object categories, and it was even possible to build highly accurate classifiers to identify from which dataset images were drawn. Taken together, these results imply that none of the data sets are unbiased samples of the underlying object categories.

Third, there is plenty of evidence that the firing of neurons in high-level visual cortex depends on more than just the scene being viewed. Neuronal activity in visual cortex is known to depend strongly on task demands, attentional allocation [19,20], and even expectation [21••,22]. Understanding how population representations are modulated by these 'external' factors, and how downstream neurons read-out information in the face of these factors is important.

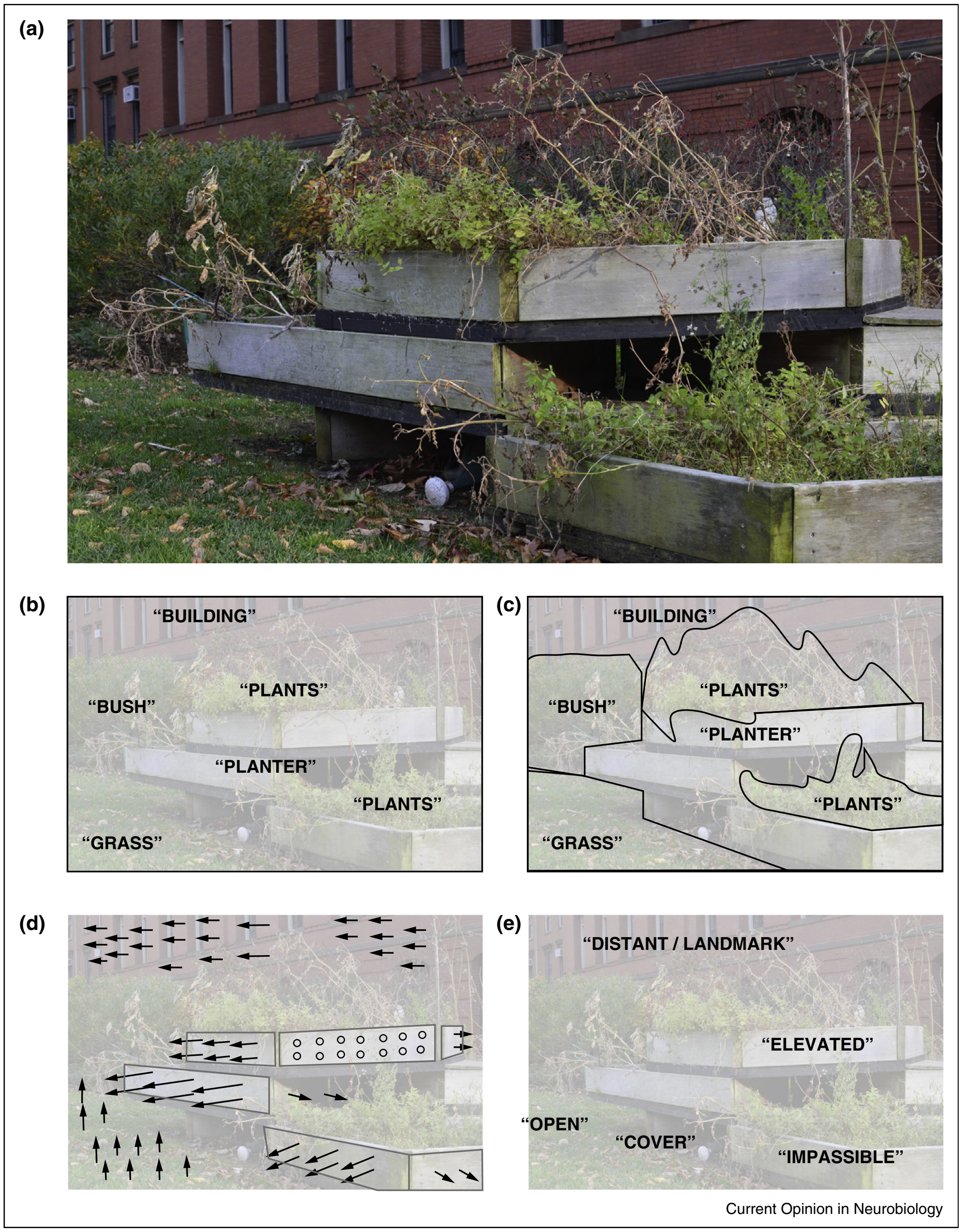
## Is object recognition even the correct framing problem?

As an operational definition of 'higher-level vision', object recognition has helped us gain a useful foothold in understanding some of the basic properties of high-level visual cortex. It is, however, a limited operational framing of the high-level vision problem. A deeper problem is that object recognition and categorization are only a small slice of our visual systems can and must do. Consider the scene in Figure 2. While it true that the scene contains identifiable objects, a purely label-based readout of the scene (Figure 2b) is just a shadow of the wealth of information available in the scene (Figure 2d–e). It is true that animals must identify certain classes of objects — food, mates, and predators, to name a few. There are other important parameters of any object beyond its identity: How big is it? How close or far? What is it made of? Moreover, there are a wide range of tasks that defy the simple 'object recognition' framing: How do we decide whether a particular path will be passable or impassable? How do we decide the orientation of surfaces so that we can know if we must climb up them or step over them? How do we decide whether a place is one we have been in before? How do we make judgments about the material properties of surfaces that are not easily labeled?

This dissonance is thrown into even starker relief when we think about vision in mammals other than primates. While object recognition per se may be of obvious relevance to primates, which evolved to seek out fruit in cluttered arboreal environments, a central focus on object recognition is less obviously important in other mammals. Recently, we and others have shown that rodents can be trained to perform complex visual recognition tasks [23–25], demonstrating that they possess unexpectedly sophisticated visual abilities. However, it seems unlikely that object recognition and discrimination are the most important higher-level visual functions in rodents. Rodents possess impressive abilities to navigate in complex environments, foraging over large ranges in a fast-changing lighting environment of the twilight hours. Navigation presents many of the same fundamental demands for view and illumination tolerance as object recognition, yet the visual orienting and scene parsing required for navigation do not obviously reduce to series of discrete object recognition tasks.

All of the above raises an important question for our field: Should we reframe our approach to tackling the vision problem? Work on shape coding in V4 and IT [26–28] begins to explore interesting questions beyond object recognition. However, much of this work is still confined to stimuli explicitly designed to test hypotheses about shape coding (isolated fragments on a blank background, clear bounding contours, etc.); such models have not yet been applied successfully to predict responses to real world scenes.

Figure 2



Current Opinion in Neurobiology

Is object recognition the correct framing problem for high-level vision? **(a)** While much research has focused on object recognition as a central task in vision, many real world scenes are only poorly described by object labels. While it is certainly possible to apply objects labels to elements in this scene **(b)**, such labels provide an extremely impoverished description of the scene, and it is unclear whether some of the labels (e.g. 'planter') are valid for animals besides humans). **(c)** A segmentation-based description of the scene is better, but still represents a shadow of the total information content of the scene. We are easily able to extra a wealth of information from this scene, even for objects that are problematic to label. This additional information includes 3D information, such as normal vector directions, and even more abstract task-driven information, such as a whether a portion of the scene is in the open or under cover. Such tasks may represent a more framing context for high-level vision, especially in non-human animals.

Again, when considering these 'alternative' problems in high level vision, examining the close parallels between visual neuroscience and computer vision may be instructive. Computer vision field is increasingly moving on to study 'scene understanding', moving past n-way recognition tasks such as the Caltech 101 and Pascal VOC, and moving towards problems that require extracting dense labeling of pixels [29,30], intra-image structural relationships [31], and even local 3D structure [32]. Work is also being done to explore the differential demands of visual mapping and spatial localization, sometimes even with direct analog to biology [33,34•].

While much work in computer science is more focused on application and theory than on biology, the field nonetheless serves as a fertile ground for visual neuroscience to explore what guiding questions neuroscientists ask. Richard Feynman famously declared, 'What I cannot create, I do not understand.' Given the resurgence in interest in neural network models for practical vision applications — exemplified by the rise of so-called 'deep learning' models [35] — the time has never been so ripe for direct synergy between computer vision approaches and biology. At the same time, such connections will not come without effort; biological plausibility is often a secondary consideration even in biologically inspired system. Building stronger links between computer science and visual neuroscience holds the promise to advance the state of understanding in both fields, but the ultimate test of our understanding of biological visual systems will be our ability to recreate their abilities through computational models. The key to driving our understanding of high-level vision will lie not just in searching for answers, but exploring a richer range of framing questions about what high-level vision is for.

## References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Op De Beeck H, Vogels R: **Spatial sensitivity of macaque inferior temporal neurons**. *J Comp Neurol* 2000, **426**:505-518.

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

3. Sary G, Vogels R, Orban GA: **Cue-invariant shape selectivity of macaque inferior temporal neurons**. *Science* 1993, **260**:995-997.

4. Rolls ET, Aggelopoulos NC, Zheng F: **The receptive fields of inferior temporal cortex neurons in natural scenes**. *J Neurosci* 2003, **23**:339-348.

5. Zoccolan D, Cox DD, DiCarlo JJ: **Multiple object response normalization in monkey inferotemporal cortex**. *J Neurosci* 2005, **25**:8150-8164.

6. Sugase Y, Yamane S, Ueno S, Kawano K: **Global and fine information coded by single neurons in the temporal visual cortex**. *Nature* 1999, **400**:869-873.

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

8. • Pagan M, Urban LS, Wohl MP, Rust NC: **Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information**. *Nat Neurosci* 2013, **16**:1132-1139.

This paper is an excellent example of the 'readout approach' to interrogating population responses in higher visual cortices. In this paper, the authors found that task-relevant information about visual objects was more easily decoded by a linear classifier in perirhinal cortex, as compared to inferotemporal cortex, consistent with the idea that information is progressively transformed to make it more easily decoded by a linear classifier.

9. Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P: **Distributed and overlapping representations of faces and objects in ventral temporal cortex**. *Science* 2001, **293**:2425-2430.

10. Cox DD, Savoy RL: **Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex**. *Neuroimage* 2003, **19**:261-270.

11. Kay KN, Naselaris T, Prenger RJ, Gallant JL: **Identifying natural images from human brain activity**. *Nature* 2008, **452**:352-355.

12. Naselaris T, Kay KN, Nishimoto S, Gallant JL: **Encoding and decoding in fMRI**. *Neuroimage* 2011, **56**:400-410.

13. Fei-Fei L, Fergus R, Perona P: **Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories**. *Comput Vis Image Understand* 2007, **106**:59-70.

14. Lazebnik S, Schmid C, Ponce J: **Beyond bags of features: spatial pyramid matching for recognizing natural scene categories**. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol 2*. 2006:2169-2178.

15. Bosch A, Zisserman A, Munoz X: **Representing shape with a spatial pyramid kernel**. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*. 2007:401-408.

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

17. Pinto N, Cox DD, DiCarlo JJ: **Why is real-world visual object recognition hard?** *PLoS Comput Biol* 2008, **4**.

18. •• Torralba A, Efros AA: **Unbiased look at dataset bias**. *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011:1521-1528.

This paper elegantly demonstrates the potential dangers of bias in vision benchmark datasets. In particular, the authors showed that computer vision systems trained on one computer vision benchmark set generalize less well to other data sets containing ostensibly identical object categories. Even more interestingly, the authors showed that machine learning algorithms were able to effectively learn a mapping from images to the *dataset that they came from*. This result has profound implications for the field of machine vision, but possibly even more significant ones for visual neuroscience, where less datasets tend to be more one-off and purpose-built and less widely reused.

19. Fuster JM, Jervey JP: **Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli**. *Science* 1981, **212**:952-955.

20. Reynolds JH, Chelazzi L: **Attentional modulation of visual processing**. *Annu Rev Neurosci* 2004, **27**:611-647.

21. •• Meyer T, Olson CR: **Statistical learning of visual transitions in monkey inferotemporal cortex**. *Proc Natl Acad Sci U S A* 2011, **108**:19401-19406.

This work shows surprisingly robust expectation dependent responses in inferotemporal cortex, an area of high-level visual cortex often conceptualized as a purely visual area that encodes the visual world as it occurs in the moment. Meyer and Olson showed robust responses from IT neurons to violations of learned transition probabilities in sequences of stimuli. This result implicates IT cortex in statistical learning of sequences across time, expanding the putative role of IT cortex in encoding visual experience.

22. Keller G, Bonhoeffer T, Hübener M: **Sensorimotor mismatch signals in primary visual cortex of the behaving mouse**. *Neuron* 2012, **74**:809-815.

23. Zoccolan D, Oertelt N, DiCarlo JJ, Cox DD: **A rodent model for the study of invariant visual object recognition**. *Proc Natl Acad Sci U S A* 2009, **106**:8748-8753.

24. Tafazoli S, Di Filippo A, Zoccolan D: **Transformation-tolerant object recognition in rats revealed by visual priming**. *J Neurosci* 2012, **32**:21-34.

25. Vermaercke B, Op de Beeck HP: **A multivariate approach reveals the behavioral templates underlying visual discrimination in rats**. *Curr Biol* 2012, **22**:50-55.

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

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

28. Hung C, Carlson ET, Connor CE: **Medial axis shape coding in macaque inferotemporal cortex**. *Neuron* 2012, **74**:1099-1113.

29. Li L, Socher R, Fei-Fei L: **Towards total scene understanding: classification, annotation and segmentation in an automatic framework**. *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009*. 2009:2036-2043.

30. Farabet C, Couprie C, Laurent N, LeCun Y: **Learning hierarchical features for scene labeling**. *IEEE Trans Pattern Anal Mach Intell* 2013, **35**:1915-1929.

31. Crandall D, Felzenszwalb P, Huttenlocher D: **Spatial priors for part-based recognition using statistical models**. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol 2*. 2005:10-17.

32. Saxena A, Sun M, Ng AY: **Make3d: learning 3d scene structure from a single still image**. *IEEE Trans Pattern Anal Mach Intell* 2009, **31**:824-840.

33. Milford MJ, Wyeth GF, Prasser D: **RatSLAM: a hippocampal model for simultaneous localization and mapping**. In *2004 IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04, vol 1*. 2004:403-408.

34. Milford MJ, Wiles J, Wyeth GF: **Solving navigational uncertainty**
• **using grid cells on robots**. *PLoS Comput Biol* 2010, **6**:e1000995.
Milford and colleagues present a biologically inspired implementation of visual simultaneous localization and mapping (SLAM) for robotics. This work is notable in that it builds a bridge between neuroscience and practical applications in robotics. This is important not just because it leads to applications, but because it brings hard, operationalized task demands to bear on the computational system. Crossovers like this between neuroscience and computer science/robotics hold a great deal of promise for guiding progress by providing more solid framings of relevant task demands that can be applied to neuronal systems.

35. Bengio Y, Courville A, Vincent P: *Representation Learning: A Review and New Perspectives*. 2012arXiv:1206.5538.[cs.LG].

36. Sugihara K: *Impossible Motion ''Magnet-Like Slopes''*. 2010 http://home.mims.meiji.ac.jp/∼sugihara/hobby/MLSdescription.pdf.