Opinion TRENDS in Cognitive Sciences Vol.11 No.8

Untangling invariant object recognition

James J. DiCarlo and David D. Cox

McGovern Institute for Brain Research, and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Despite tremendous variation in the appearance of visual objects, primates can recognize a multitude of objects, each in a fraction of a second, with no apparent effort. However, the brain mechanisms that enable this fundamental ability are not understood. Drawing on ideas from neurophysiology and computation, we present a graphical perspective on the key computational challenges of object recognition, and argue that the format of neuronal population representation and a property that we term 'object tangling' are central. We use this perspective to show that the primate ventral visual processing stream achieves a particularly effective solution in which single-neuron invariance is not the goal. Finally, we speculate on the key neuronal mechanisms that could enable this solution, which, if understood, would have far-reaching implications for cognitive neuroscience.

Introduction

Our daily activities rely heavily on the accurate and rapid identification of objects in our visual environment. The apparent ease of with which we recognize objects belies the magnitude of this feat: we effortlessly recognize objects from among tens of thousands of possibilities and we do so within a fraction of a second, in spite of tremendous variation in the appearance of each one. Understanding the brain mechanisms that underlie this ability would be a landmark achievement in neuroscience.

Object recognition is computationally difficult for many reasons, but the most fundamental is that any individual object can produce an infinite set of different images on the retina, due to variation in object position, scale, pose and illumination, and the presence of visual clutter (e.g. [1–5]). Indeed, although we typically see an object many times, we effectively never see the same exact image on our retina twice. Although several computational efforts have attacked this so-called 'invariance problem' (e.g. [1,3,6–12]), a robust, real-world machine solution still evades us and we lack a satisfying understanding of how the problem is solved by the brain. We believe that these two achievements will be accomplished nearly simultaneously by an approach that takes into account both the computational issues and the biological clues and constraints.

Because it is easy to get lost in the sea of previous studies and ideas, the goal of this manuscript is to clear the table, bring forth key ideas in the context of the primate brain, and pull those threads together into a coherent framework. Below, we use a graphical perspective to provide intuition about the object recognition problem, show that the primate ventral visual processing stream produces a particularly effective solution in the inferotemporal (IT) cortex, and speculate on how the ventral visual stream approaches the problem. Along the way, we argue that some approaches are only tangential to, or even distract from, understanding object recognition.

What is object recognition?

We define object recognition as the ability to accurately discriminate each named object ('identification') or set of objects ('categorization') from all other possible objects, materials, textures other visual stimuli, and to do this over a range of identity-preserving transformations of the retinal image of that object (e.g. image transformations resulting from changes in object position, distance, and pose). Of course, vision encompasses many disparate challenges that may interact with object recognition, such as material and texture recognition, object similarity estimation, object segmentation, object tracking and trajectory prediction. Exploring such possible interactions is not our goal. Instead, we aim to see how far a clear focus on the problem of object recognition will take us. We concentrate on what we believe to be the core of the brain's recognition system – the ability to rapidly report object identity or category after just a single brief glimpse of visual input (<300 ms; see Box 1) [13,14].

What computational processes must underlie object recognition?

To solve a recognition task, a subject must use some internal neuronal representation of the visual scene (population pattern of activity) to make a decision (e.g. [15,16]): is object **A** present or not? Computationally, the brain must apply a decision function [16] to divide an underlying neuronal representational space into regions where object **A** is present and regions where it is not (Figure 1b; one function for each object to be potentially reported). Because brains compute with neurons, the subject must have neurons somewhere in its nervous system – 'read-out' neurons – that can successfully report if object **A** was present [17]. Of course, there are many relevant mechanistic issues, for example, how many such neurons are involved in computing the decision, where are they in the brain, is their operation fixed or dynamically created with the task at

Corresponding authors: DiCarlo, J.J. (dicarlo@mit.edu); Cox, D.D. (cox@rowland.harvard.edu).

Available online 16 July 2007.

 1364-6613/$ – see front matter doi:10.1016/j.tics.2007.06.010

Box 1. Frequently asked questions

Feed-forward versus feedback?
We do not mean to imply that all recognition is a result of feed-forward mechanisms (e.g. Figure 2 in the main text). However, the psychophysical and physiological data suggest that transformation to support recognition 'in a glimpse' takes <300 ms from time of stimulus onset [13,14,19]. Even the earliest IT spikes (~125 ms latency) can already support robust recognition [19]. This places serious constraints on the types of feedback that might be involved. For example, such data argue against the possibility that an initial IT (or higher) representation is created (e.g. for prior model selection or feature attention set) and information is fed back down the hierarchy to V1 and then back up the hierarchy, before a good IT representation occurs. However, these data do not preclude feedback within cortical areas or possibly between neighboring cortical areas (e.g. V2 to V1; see Figure 2 in the main text). They also do not mean that priors cannot be used (see below). Also, not all recognition occurs in a glimpse and does involve top-down feedback processes (e.g. attentional shifts).

What about visual clutter?
We have focused on what happens to an object's neuronal image as a result of identity-preserving transformations that are easily parameterized (e.g. pose, position, size). However, one major source of real-world image variation is clutter – backgrounds and other potentially occluding objects, which might, in turn, cast shadows and other variations in object illumination. Although not as obvious, this variation results in the same types of object manifolds as already described (relatively low-dimensional, continuous and highly curved in retinal image space). Thus, the essence of the problem is still the same in that it is another source of identity-preserving variation that tends to tangle the object manifolds.

What about using priors to improve recognition?
In severely occluding clutter or severely foreshortened views, different objects can produce the same retinal image (one situation where object manifolds touch). Such cases are fundamentally ambiguous and can only be 'solved' with prior assumptions (e.g. [11,65]). This might not be done in a glimpse and might require large-scale feedback (see above). However, this does not mean that priors cannot be involved in recognition in a glimpse. Indeed, we think priors are involved in an implicit, largely feed-forward way in that, at each ventral stream processing stage, neuronal tuning functions 'looking' at the previous stage over-represent features and feature contingencies that are most often encountered in the world, rather than, say, white noise [57,59] (see text). That is, we argue that real-world primate object recognition uses implicit priors all of the time, and top-down priors some of the time, depending on task demands.

hand, and how do they code choices in their spikes? However, these are not the central computational issues of object recognition. The central issues are: what is the format of the representation used to support the decision (the substrate on which the decision functions directly operate); and what kinds of decision functions (i.e. read-out tools) are applied to that representation?

These central computational issues are two sides of the same coin. For example, one can view object recognition as the problem of finding very complex decision functions (highly non-linear) that operate on the retinal image representation. Alternatively, one can view it as the problem of finding operations that progressively transform that retinal representation into a new form of representation, followed by the application of relatively simple decision functions (e.g. linear classifiers [18]). From a computational perspective, the difference is largely terminology, but we and others (e.g. [16,19]) argue that the latter viewpoint is more productive because it starts to take the problem apart in a way that is consistent with the architecture and response properties of the ventral visual stream, and because such simple decision functions are easily implemented in a single, biologically plausible neuronal processing step (a thresholded sum over weighted synapses). This view also meshes well with conventional pattern recognition wisdom – choice of representation is often more important than the 'strength' of the classifier used. As shown below, a variety of recognition tasks can be solved in IT cortex population responses using simple, linear classifiers [19], suggesting that our focus on such operations is not unreasonable. Finally, even from this viewpoint, one is completely free to consider the possibility that the algorithms that implement 'representation' are not different from those applied during 'decision'. Thus, with little loss of generality, below we treat object recognition fundamentally as a problem of data representation and re-representation, and we use simple decision functions (linear classifiers) to examine those representations.

Why is object recognition hard? Object manifold tangling

Object recognition is hard because useful forms of visual representation are hard to build. A major impediment to understanding such representations arises from the fact that vision operates in high-dimensional space. Our eyes fixate the world in ~300 ms intervals before moving on to a new location. During each brief glimpse, a visual image is projected into the eye, transduced by ~100 million retinal photoreceptors and conveyed to the brain in the spiking activity pattern of ~1 million retinal ganglion cells. Such a representation can be conceptualized as a high-dimensional extension of a simple three-dimensional Cartesian space in which each axis of the space is the response of one retinal ganglion cell (e.g. [20,21]) (Figure 1). Even ignoring temporal information and measuring the response of each neuron to each glimpse as its mean spiking rate, each image projected into the eye is one point in an ~1 million dimensional retinal ganglion cell representation (Box 1).

To gain intuition about high-dimensional visual representations, note that, within this immense retinal representation space, different encounters with the same physical object lie in contiguous regions. For example, consider one glimpse of a particular face. That single glimpse of that face, in exactly that position, scale, pose, lighting and background, produces just one pattern of activity on your retina – it is just one point in retinal image space (internal neuronal 'noise' would introduce a small amount of variation in each point, but is ignored here because it is not fundamental to our arguments). Now imagine all the possible retinal images that face could ever produce (e.g. due to changes in its pose, position and size) and the corresponding set of points in retinal image space. That set of potential data points arises from a continuous, low-dimensional, curved surface inside the retinal image space called an object 'manifold' (Figure 1a) [3,20,21]. Different objects have different manifolds (Figure 1b–d).

Given this framework, we consider a simple world with just two possible objects (Joe and Sam, see Figure 1) to

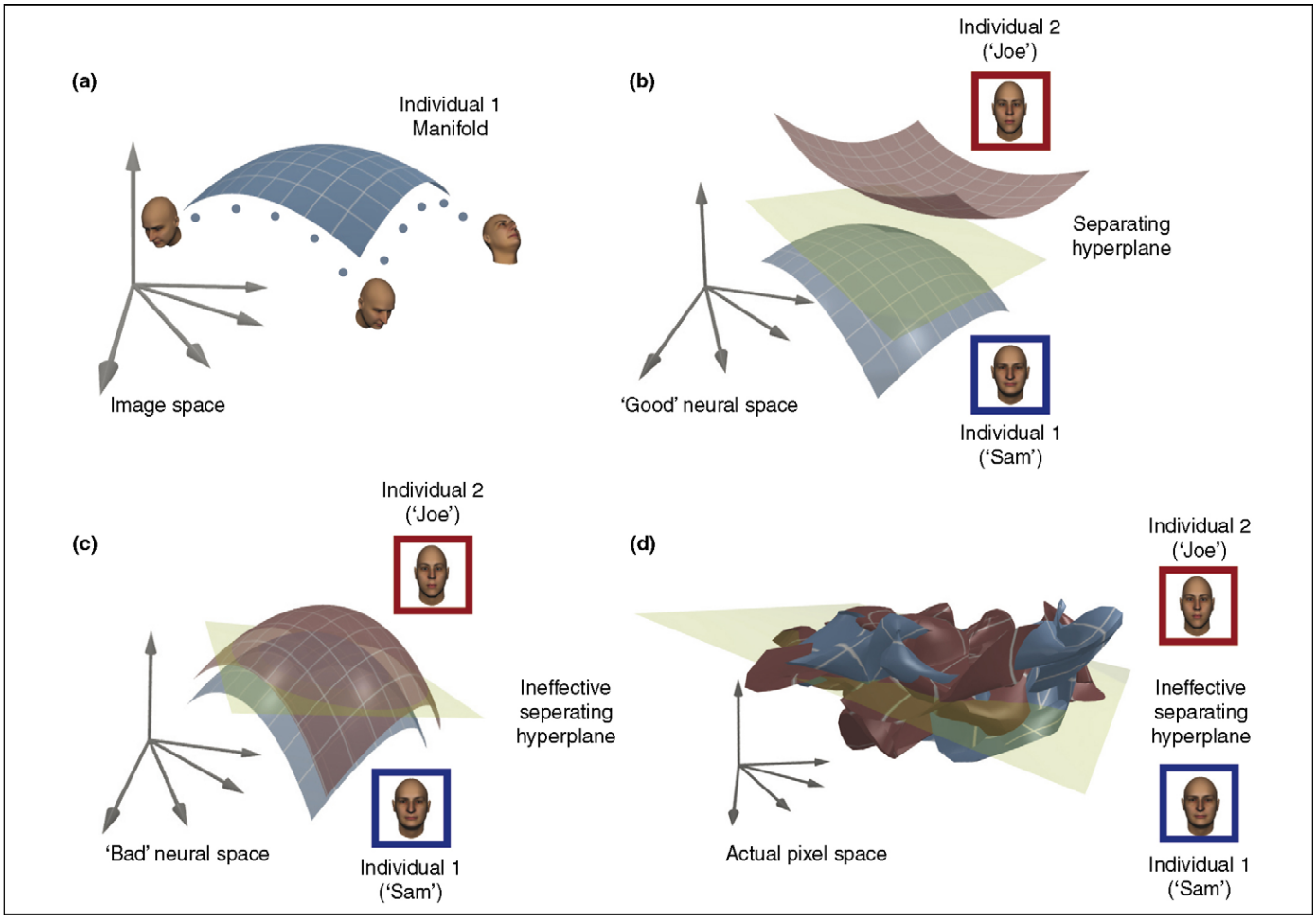

Figure 1. Illustration of object tangling. In a neuronal population space, each cardinal axis is one neuron's activity (e.g. firing rate over an ~200 ms interval) and the dimensionality of the space is equal to the number of neurons. Although such high-dimensional spaces cannot be visualized, the three-dimensional views portrayed here provide fundamental insight. **(a)** A given image of a single object (here, a particular face) is one point in retinal image space. As the face's pose is varied, the point travels along curved paths in the space, and all combinations of left/right and up/down pose (two degrees of freedom) lie on a two-dimensional surface, called the object manifold (in blue). Although only two degrees of freedom are shown for clarity, the same idea applies when other identity-preserving transformations (e.g. size, position) are applied. **(b)** The manifolds of two objects (two faces, red and blue) are shown in a common neuronal population space. In this case, a decision (hyper-) plane can be drawn cleanly between them. If the world only consisted of this set of images, this neuronal representation would be 'good' for supporting visual recognition. **(c)** In this case, the two object manifolds are intertwined, or tangled. A decision plane can no longer separate the manifolds, no matter how it is tipped or translated. **(d)** Pixel (retina-like) manifolds generated from actual models of faces (14,400-dimensional data; 120 × 120 images) for two face objects were generated from mild variation in their pose, position, scale and lighting (for clarity, only the pose-induced portion of the manifold is displayed). The three-dimensional display axes were chosen to be the projections that best separate identity, pose azimuth and pose elevation. Even though this simple example only exercises a fraction of typical real-world variation, the object manifolds are hopelessly tangled. Although the manifolds appear to cross in this three-dimensional projection, they do not cross in the high-dimensional space in which they live.

graphically show the difference between a 'good' and 'bad' representation for directly supporting object recognition. The representation in Figure 1b is good: it is easy to determine if Joe is present, in spite of pose variation, by simply placing the linear decision function (i.e. a hyperplane) between Joe's manifold and the other potential images in the visual world (just images of Sam in this case, but see Figure I in Box 2). By contrast, the representation in Figure 1c is bad: the object manifolds are tangled, such that it is impossible to reliably separate Joe from the rest of the visual world with a linear decision function. Figure 1d shows that this problem is not academic – the manifolds of two real-world objects are hopelessly tangled together in the retinal representation.

Note, however, that the two manifolds in Figure 1c,d do not cross or superimpose – they are like two sheets of paper crumpled together. This means that, although the retinal representation cannot directly support recognition, it implicitly contains the information to distinguish which of the two individuals was seen. We argue that this describes the computational crux of 'everyday' recognition: the problem is typically not a lack of information or noisy information, but that the information is badly formatted in the retinal representation – it is tangled (but also see Box 1). Although Figure 1 shows only two objects, the same arguments apply when more objects are in the world of possible objects – it just makes the problem harder, but for exactly the same reasons.

One way of viewing the overarching goal of the brain's object recognition machinery, then, is as a transformation from visual representations that are easy to build (e.g. center-surround filters in the retina), but are not easily decoded (as in Figure 1c,d), into representations that we do not yet know how to build (e.g. representations in IT), but are easily decoded (e.g. Figure 1b). Although the idea of representational transformation has been stated under

Box 2. The power, and challenge, of high-dimensional representations

Although Figures 1d and 3a,b in the main text shows that object manifolds are less tangled in some neuronal population representations than others, they also necessarily hide the full complexity and power of high-dimensional representations. One false impression these figures might create is that the untangling perspective only applies when the world contains just two possible objects. However, these three-dimensional pictures are just projections of a high-dimensional space in which it is easy for a large number of possible object manifolds to be represented such that they are mutually separable yet still maintain correspondence of transformation variables (e.g. pose, position and size). Figure I shows an example with five hypothetical object manifolds. All panels show three-dimensional projections (views) of exactly the same seven-dimensional representation. Projections exist in which a hyperplane cleanly separates each object manifold from the others (Figure Ic,d). At the same time, the manifolds are coordinated in that projections also exist that are useful for discriminating the pose of the object, for example (Figure Ib). It may seem almost magical that just looking at the same representation from a different perspective can bring each of the manifolds out to one side of the space. Indeed, this would be impossible in three dimensions, but it is straightforward in many dimensions.

Another false impression that Figures 1 and 3 in the main text might create is that a linear decision plane approach would necessarily result in many large recognition errors (over-generalization), because the decision plane accepts everything on one side of it as being 'object present', including points that might be arbitrarily far from the object manifold. However, this is not a serious limitation because the full volume of a real neuronal representation space usually cannot be reached (i.e. not all patterns of population response are possible). For example, because neurons have minimum and maximum response values (e.g. firing rates), the reachable space is limited to a hyper-rectangle. Dependencies between the neural response functions (e.g. due to common inputs) further limit the reachable space. Indeed, because the simulated IT neurons in Figure 3b are Gaussian functions, no stimulus exists that can produce a population response that is very far out on the 'Joe detected' side of Joe's object manifold. Figure Ie explores this idea further. Nevertheless, such spaces can support extremely rich representations because the reachable hypersurface has many degrees of freedom and a massive surface area.

In sum, one can imagine a good high-dimensional object representation where all object manifolds lie near the 'edges' of the reachable part of the population space (Figure Ie) and are mutually orthogonal (Figure Ia–d). The population of IT cells suggested by existing data (see Figure 3b) is consistent with this viewpoint.

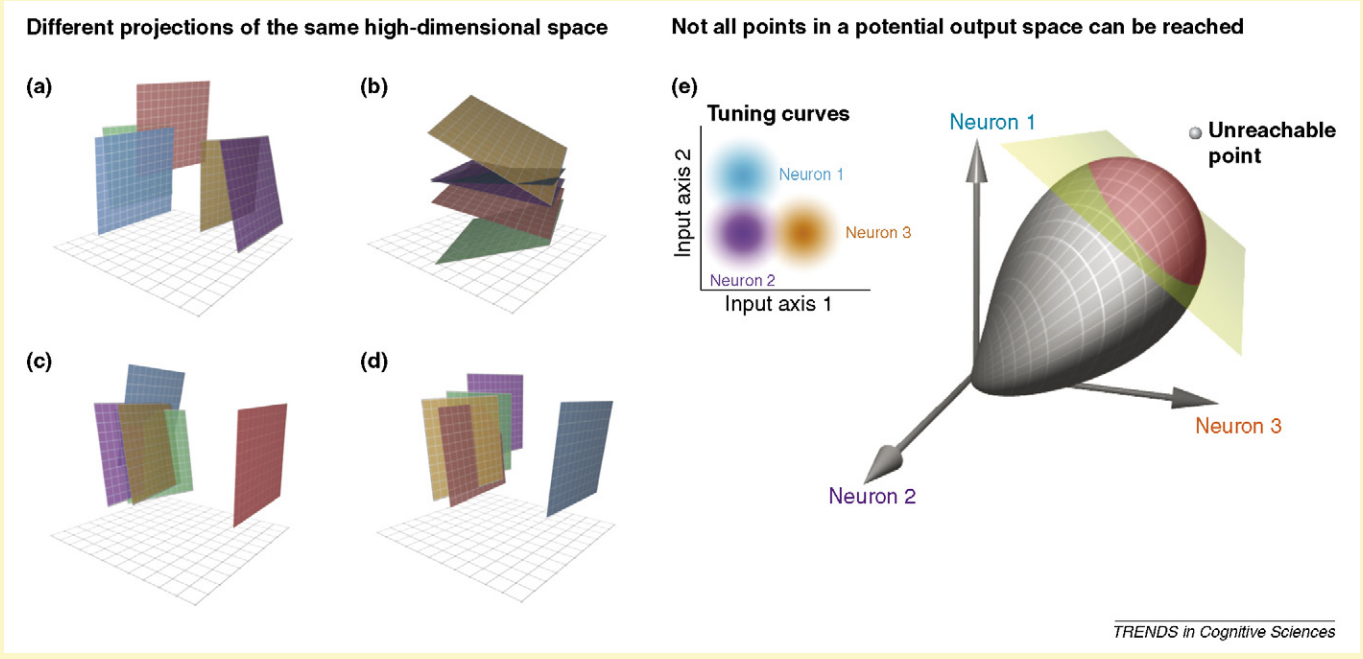

Figure I. The challenges of thinking in high-dimensional spaces. **(a–d)** Three-dimensional projections of the same exact seven-dimensional population representation. Each colored plane is one of five perfectly flat object manifolds (analogous to the manifolds in Figures 1 and 3 in the main text). In the projection in (a), it is difficult to see how each manifold could be separated from the others using decision hyperplanes; it might be possible to separate the purple manifold from the rest, but it does not seem possible to cleanly separate the red or blue manifolds from the others. However, panels (c and d) show projections (views) of the same representation in which the red and blue object manifolds are easily separated from the rest. Likewise, whereas the degrees of freedom of the manifolds (analogous to 'pose' in Figures 1 and 3) do not line up in (a,c and d), there nonetheless exist projections where they do line up (b). Such a projection would be useful for reading out the pose of the object, independent of identity. Panel **(e)** illustrates that not all potential locations of a population representation are reachable. It shows a population space spanned by just three simulated neurons, each with Gaussian tuning in a two-dimensional input space (inset). The responses of these three neurons are plotted to the right. Because the neuronal response functions are overlapping (non-independent), it is only possible to reach points in the output space that are on the gray surface shown here. If one uses a hyperplane (shown in green) to demarcate a region corresponding to an object (shown in red), then one need not worry about falsely including extraneous over-generalized points on the 'object-detected' side of the plane (e.g. the point labeled unreachable point), because the neuronal population can never give that pattern of response.

many guises (e.g. 2 1/2D sketch and feature selection [18,22,23]), we argue below that the untangling perspective goes farther, suggesting the kinds of transformations the ventral visual system should perform. However, first we look at the primate ventral visual stream from this untangling perspective.

The ventral visual stream transformation untangles object manifolds

In humans and other primates, information processing to support visual recognition takes place along the ventral visual stream (for reviews, see [5,24,25]). We, and others (e.g. [1,26]), consider this stream to be a progressive series

of visual re-representations, from V1 to V2 to V4 to IT cortex (Figure 2). Beginning with the studies of Gross [27], a wealth of work has shown that single neurons at the highest level of the monkey ventral visual stream – the IT cortex – display spiking responses that are probably useful for object recognition. Specifically, many individual IT neurons respond selectively to particular classes of objects, such as faces or other complex shapes, yet show some tolerance to changes in object position, size, pose and illumination, and low-level shape cues. (Also see e.g. Ref. [28] for recent related results in humans.)

How can the responses of individual ventral stream neurons provide insight into object manifold untangling in the brain? To approach this, we have focused on characterizing the initial wave of neuronal population 'images' that are successively produced along the ventral visual stream as the retinal image is transformed and re-represented on its way to the IT cortex (Figure 2). For example, we and our collaborators recently found that simple linear classifiers can rapidly (within <300 ms of image onset) and accurately decide the category of an object from the firing rates of an IT population of ~200 neurons, despite variation in object position and size [19]. It is important to note that using 'stronger' (e.g. non-linear) classifiers did not substantially improve recognition performance and the same classifiers fail when applied to a simulated V1 population of equal size [19]. This shows that performance is not a result of the classifiers themselves, but the powerful form of visual representation conveyed by the IT cortex. Thus, compared with early visual representations, object manifolds are less tangled in the IT population representation.

To show this untangling graphically, Figure 3 illustrates the manifolds of the faces of Sam and Joe from Figure 1d (retina-like representation) re-represented in the V1 and IT cortical population spaces. To generate these, we took populations of simulated V1-like response functions (e.g. Refs [29,30]) and IT-like response functions (e.g. Refs [31,32]), and applied them to all the images of Joe and Sam. This reveals that the V1 representation, like the retinal representation, still contains highly curved, tangled object manifolds (Figure 3a), whereas the same object manifolds are flattened and untangled in the IT representation (Figure 3b). Thus, from the point of view of downstream decision neurons, the retinal and V1 representations are not in a good format to separate Joe from the rest of the world, whereas the IT representation is. In sum, the experimental evidence suggests that the ventral stream transformation (culminating in IT) solves object recognition by untangling object manifolds. For each visual image striking the eye, this total transformation happens progressively (i.e. stepwise

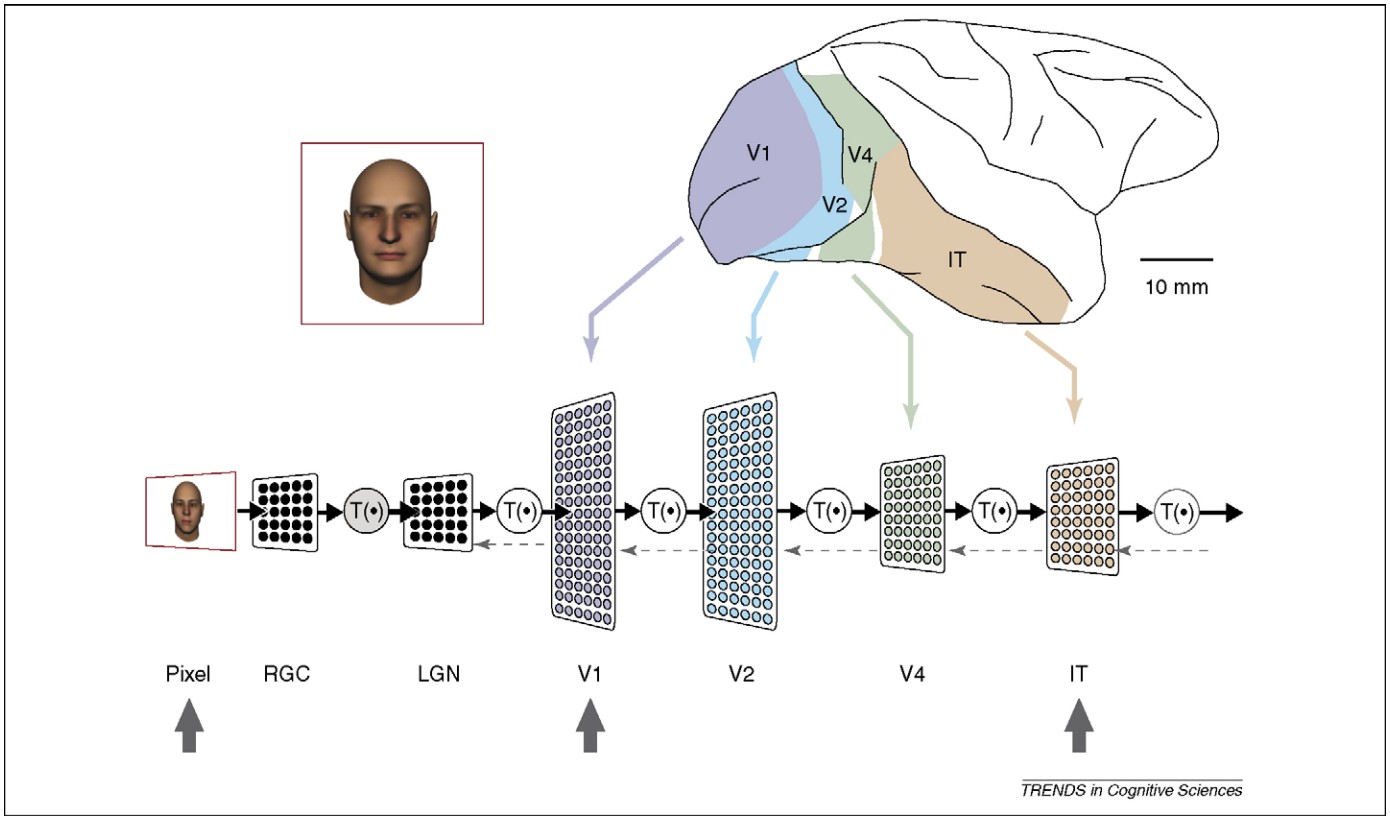

Figure 2. Neuronal populations along the ventral visual processing stream. The rhesus monkey is currently our best model of the human visual system. Like humans, monkeys have high visual acuity, rely heavily on vision (~50% of macaque neocortex is devoted to vision) and easily perform visual recognition tasks. Moreover, the monkey visual areas have been mapped and are hierarchically organized [26], and the ventral visual stream is known to be critical for complex object discrimination (colored areas, see text). We show a lateral schematic of a rhesus monkey brain (adapted from Ref. [26]). We conceptualize each stage of the ventral stream as a new population representation. The lower panels schematically illustrate these populations in early visual areas and at successively higher stages along the ventral visual stream – their relative size loosely reflects their relative output dimensionality (approximate number of feed-forward projection neurons). A given pattern of photons from the world (here, a face) is transduced into neuronal activity at the retina and is progressively and rapidly transformed and re-represented in each population, perhaps by a common transformation (T). Solid arrows indicate the direction of visual information flow based on neuronal latency (~100 ms latency in IT), but this does not preclude fast feedback both within and between areas (dashed arrows, see Box 1). The gray arrows across the bottom indicate the population representations for the retina, V1 and IT, which are considered in Figures 1d and 3a,b, respectively. RGC, retinal ganglion cells; LGN, lateral geniculate nucleus.

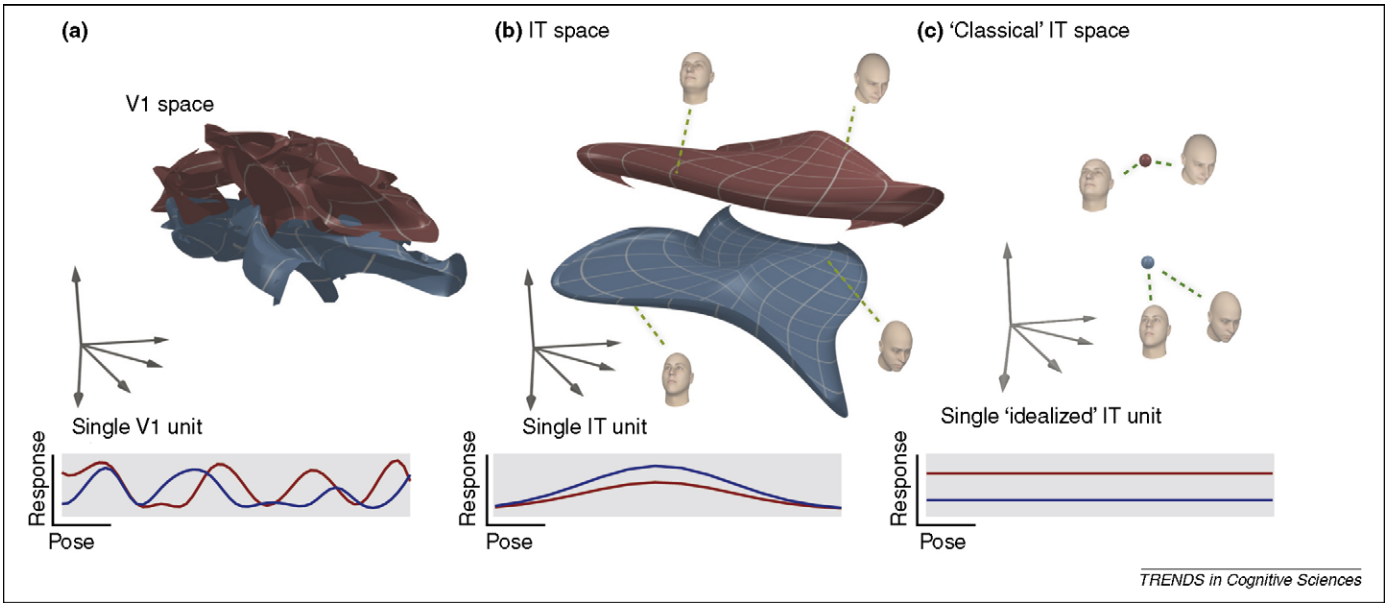

Figure 3. Untangling object manifolds along the ventral visual stream. As visual information progresses through the ventral visual pathway, it is progressively re-represented in each visual area and becomes better and better at directly supporting object recognition. **(a)** A population of 500 V1 neurons was simulated as a bank of Gabor filters with firing thresholds. Display axes in this 500-dimensional population space were chosen to maximally separate two face stimuli undergoing a range of identity-preserving transformations (pose, size, position and lighting direction), as in Figure 1. Manifolds are shown for the two objects (red and blue) undergoing two-axis pose variation (azimuth and elevation). As with the retina-like space shown in Figure 1c, object manifolds corresponding to the two objects are hopelessly tangled together. Below, the responses of an example single unit are shown in response to the two faces undergoing one axis of pose variation. **(b)** By contrast, a population of simulated IT neurons gives rise to object manifolds that are easily separated. 500 IT neurons were simulated with broad (but not flat) unimodal Gaussian tuning with respect to identity-preserving transformations and with varying levels of preference for one or the other face, analogous to what is observed in single unit recording in IT. In addition to being able to separate object manifolds corresponding to different identities, such a representation also allows one to recover information about object pose. The lines going through the two manifolds show that the manifolds are coordinated – they are lined up in such a way that multiple orthogonal attributes of the object can be extracted using the same representation. It is important to note that, in contrast to the V1 simulation, we do not yet know how to generate single unit responses like this from real images. **(c)** A textbook idealized IT representation also produces object manifolds that are easy to separate from one another in terms of identity. Here, IT neurons were simulated with idealized, perfectly invariant receptive fields. However, although this representation may be good for recovering identity information, it 'collapses' all other information about the images.

along the cortical stages), but rapidly (i.e. <100 ms from V1 to IT, ~20 ms per cortical stage). But what is this transformation? That is, how does the ventral stream do this?

How does the ventral visual stream untangle object manifolds?

We do not yet know the answer to this question. Hubel and Wiesel's [30] observation that visual cortex complex cells can pool over simple cells to build tolerance to identity-preserving transformations (especially position) has been computationally implemented and extended to higher cortical levels, including the IT [1,12,33]. However, beyond this early insight, systems neuroscience has not provided a breakthrough.

Some neurophysiological effort has focused on characterizing IT neuronal tolerance to identity-preserving transformations (e.g. Refs [31,32,34–38]), which is central to object tangling. However, much more effort has been aimed at understanding the effects of behavioral states, for example, task and attention (e.g. Refs [39–45]). Although important, these studies sidestep the untangling problem, because such effects can be measured without understanding the format of representation.

Substantial effort has also recently been aimed at understanding the features or shape dimensions of visual images to which V4 and IT neurons are most sensitive (e.g. Refs [25,46–51]). Such studies are important for defining the feature complexity of ventral stream neuronal tuning, which is related to manifold untangling (because 'object' or feature conjunction manifolds are what must be untangled). Ongoing, ambitious approaches to understanding the response functions of individual neurons (i.e. the non-linear operators on the visual image) would, if successful, lead to an implicit understanding of object representation. However, given the enormity of this task, it is not surprising that progress has been slow.

The object untangling perspective leads to a complementary but qualitatively different approach. First, it shifts thinking away from single IT neuron response properties [17] – which is akin to studying feathers to understand flight [22] – toward thinking about ideal population representations, with the computational goals of the task clearly considered (see Figure 3b versus 3c) [52]. Second, it suggests the immediate goal of determining how well each ventral stream neuronal representation has untangled object manifolds and shows how to quantitatively measure untangling (see linear classifiers above, Figure 1). Third, this perspective points to better ways to compare computational models to neuronal data: whereas model predictions at the single-unit level are typically grossly under-constrained, population-level comparisons might be more meaningful (e.g. the predicted degree of untangling at each ventral stream stage). Fourth, it suggests a clear focus on the causes of tangling – identity-preserving transformations – rather than the continuing primary focus on 'shape' or 'features'. Indeed, because we do not understand the dimensions of 'shape', we speculate that computational approaches that focus on building

tolerance across identity-preserving transformations while simply preserving sensitivity to other real-world image variation will be vastly more tractable. Finally, this perspective steers experimental effort toward testing hypothetical mechanisms that might underlie untangling (e.g. Refs [53,54]), and directs complementary computational effort toward finding new biologically plausible algorithms that progressively untangle object manifolds (e.g. Refs [1,7]; see below).

Flattened object manifolds are a good solution

Figure 3 suggests a strategy for building good object representations: if the goal is to untangle manifolds corresponding to different objects, then we seek transformations that 'flatten' these manifolds, while keeping them separate (i.e. preserving 'shape' information). This perspective is partly a restatement of the problem of invariant object recognition, but not an entirely obvious one. For example, the textbook conception of IT cortex suggests a different set of goals for each IT neuron: high shape selectivity and high 'invariance' to identity-preserving image transformations. To illustrate how object manifold untangling gives a fresh perspective, Figure 3b,c shows two simulated IT populations that have both successfully untangled object identity, but that have very different single-unit response properties. In Figure 3c, each single unit has somehow met the textbook ideal of being selective for object identity, yet invariant to identity-preserving transformations. At the IT population level, this results in the untangling of object manifolds by 'collapsing' each manifold to a single point. By comparison, in Figure 3b, every single IT unit has good sensitivity to object identity, but only limited tolerance to object transformation (e.g. position, scale, view) and, by textbook standards, seems less than ideal. However, at the population level, this also results in untangled object manifolds, but in a way that has 'flattened' and coordinated, rather than discarded, information about the transformation variables (e.g. pose, position and scale). This suggests that a properly untangled IT representation (e.g. Figure 3b, Box 2) can not only directly support object recognition, but also support tasks such as pose, position and size estimation, as previously suggested by theorists (e.g. [3,19]). Indeed, real IT neurons are not, for example, position and size invariant, in that they have limited spatial receptive fields [32,36]. It is now easy to see that this 'limitation' is an advantage.

Ways the brain might flatten object manifolds

Although object manifold flattening might be partly accomplished by hard-wired transformations (e.g. Refs [1,33]), one could also learn the structure of manifolds from the statistics of natural images (e.g. Refs [20,21]), potentially allowing them to be flattened. Although most previous manifold learning efforts have emphasized learning structure in the ambient pixel/retina space in one step, we impose no such requirement. In particular, the transformations need only flatten object manifolds progressively, in a series of rapidly executed, successive steps (consistent with physiological data along the ventral stream [5]). Progressive flattening is a matter of both emphasis and substance: there is no need to swallow the entire problem whole – flattening at local, small scales can ultimately produce flattening at a global scale. Indeed, the manifold untangling perspective makes sense at a variety of scales – V1 neurons in a local neighborhood only 'see' the world through a small aperture (they cannot see whole objects), but they can perform flattening operations with respect to their inputs; V2 can do the same on its V1 inputs and so on. Thus, we believe that the most fruitful computational algorithms will be those that a visual system (natural or artificial) could apply locally and iteratively at each cortical processing stage (e.g. Refs [1,12,33,55]) in a largely unsupervised manner (e.g. Ref. [56]), and that achieve some local object manifold flattening. Even though no single cortical stage or local ensemble within a stage would 'understand' its global role, we imagine the end result to be globally flattened, coordinated object manifolds with preserved shape selectivity (Figure 3b, Box 2).

Three computational ideas that are consistent with physiology might, together, enable manifold flattening. First, the visual system projects incoming information into an even higher dimensional overcomplete space (e.g. ~100 times more V1 neurons than retinal ganglion neurons) (Figure 2). This 'spreads out' the data into a much larger space. The additional constraint of response sparseness can reduce the size of the subspace that any given incoming visual image 'lives' in and thus makes it easier to find projections where object manifolds are flat and separable (see Refs [57,58]). A second related idea is that, at each processing stage, neuronal resources (i.e. neuronal tuning functions on inputs from the previous stage) are allocated in a way that matches the distribution of visual information encountered in the real world (e.g. Refs [59,60]). This would increase the effective over-completeness of visual representations of real-world objects (and thus help flatten object manifolds). Indeed, a variety of biologically plausible algorithms recently developed in other contexts (e.g. [58,61]) might have a yet to be discovered role in achieving coordinated flattening within local neuronal populations. For example, divisive normalization is a powerful non-linearity that can literally 'bend' representational spaces.

A third, potentially key, idea is that time implicitly supervises manifold flattening. Several theorists have noticed that the temporal evolution of a retinal image provides clues to learning which image changes are identity-preserving transformations and which are not [6,7,62–64]. In the language of object tangling, this is equivalent to saying that temporal image evolution spells out the degrees of freedom of object manifolds. The ventral stream might use this temporal evolution to achieve progressive flattening of object manifolds across neuronal processing stages. Indeed, recent studies in our laboratory [53] and others [54] have begun to connect this computational idea with biological vision, showing that invariant object recognition can be predictably manipulated by the temporal statistics of the environment.

Much work must now be done to continue down this path of discovery – we still do not understand object recognition. Nevertheless, it is a very exciting time as there is a rapid blurring of lines between traditionally separate disciplines, and we hope that the perspective

presented here will galvanize efforts on one of the most exciting problems in cognitive and systems neuroscience.

Acknowledgements

We would like to thank N Kanwisher, N Li, N Majaj, N Rust, J Tenenbaum and three anonymous referees for helpful comments on earlier versions of this manuscript. Support was provided by The National Eye Institute (NIH-R01-EY014970), The Pew Charitable Trusts (PEW UCSF 2893sc) and The McKnight Foundation.

References

1 Riesenhuber, M. and Poggio, T. (1999) Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025
2 Ullman, S. (1996) *High Level Vision,* MIT Press
3 Edelman, S. (1999) *Representation and Recognition in Vision,* MIT Press
4 Ashbridge, E. and Perrett, D.I. (1998) Generalizing across object oreintation and size. In *Perceptual Constancy* (Walsh, V. and Kulikowski, J., eds), pp. 192–209, Cambridge University Press
5 Rolls, E.T. (2000) Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron* 27, 205–218
6 Ullman, S. and Soloviev, S. (1999) Computation of pattern invariance in brain-like structures. *Neural Netw.* 12, 1021–1036
7 Wallis, G. and Rolls, E.T. (1997) Invariant face and object recognition in the visual system. *Prog. Neurobiol.* 51, 167–194
8 Biederman, I. (1987) Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147
9 Olshausen, B.A. *et al.* (1993) A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.* 13, 4700–4719
10 Arathorn, D. (2002) *Map-seeking Circuits in Visual Cognition,* Stanford University Press
11 Yuille, A. and Kersten, D. (2006) Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* 10, 301–308
12 Serre, T. *et al.* (2007) Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 411–426
13 Thorpe, S. *et al.* (1996) Speed of processing in the human visual system. *Nature* 381, 520–522
14 Potter, M.C. (1976) Short-term conceptual memory for pictures. *J. Exp. Psychol. [Hum Learn].* 2, 509–522
15 Ashby, F.G. and Gott, R.E. (1988) Decision rules in the perception and categorization of multidimensional stimuli. *J. Exp. Psychol. Learn. Mem. Cogn.* 14, 33–53
16 Johnson, K.O. (1980) Sensory discrimination: decision process. *J. Neurophysiol.* 43, 1771–1792
17 Barlow, H. (1995) The neuron doctrine in perception. In *The Cognitive Neurosciences* (Gazzaniga, M.S., ed.), pp. 415–435, MIT Press
18 Duda, R.O. *et al.* (2001) *Pattern Classification,* Wiley Interscience
19 Hung, C.P. *et al.* (2005) Fast readout of object identity from macaque inferior temporal cortex. *Science* 310, 863–866
20 Tenenbaum, J.B. *et al.* (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323
21 Roweis, S.T. and Saul, L.K. (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* 290, 2323–2326
22 Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information,* Henry Holt & Company
23 Johnson, K. *et al.* (1995) *Neural Mechanisms of Tactile Form Recognition,* MIT Press
24 Logothetis, N.K. and Sheinberg, D.L. (1996) Visual object recognition. *Annu. Rev. Neurosci.* 19, 577–621
25 Tanaka, K. (1996) Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19, 109–139
26 Felleman, D.J. and Van Essen, D.C. (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47
27 Gross, C.G. *et al.* (1972) Visual properties of neurons in inferotemporal cortex of the Macaque. *J. Neurophysiol.* 35, 96–111
28 Quiroga, R.Q. *et al.* (2005) Invariant visual representation by single neurons in the human brain. *Nature* 435, 1102–1107
29 Ringach, D.L. (2002) Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J. Neurophysiol.* 88, 455–463
30 Hubel, D.H. and Wiesel, T.N. (1977) Ferrier lecture. Functional architecture of macaque monkey visual cortex. *Proc. R. Soc. Lond. B. Biol. Sci.* 198, 1–59
31 Logothetis, N.K. *et al.* (1995) Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* 5, 552–563
32 Op de Beeck, H. and Vogels, R. (2000) Spatial sensitivity of macaque inferior temporal neurons. *J. Comp. Neurol.* 426, 505–518
33 Fukushima, K. (1980) Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193–202
34 Tovée, M.J. *et al.* (1994) Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert monkey. *J. Neurophysiol.* 72, 1049–1060
35 Ito, M. *et al.* (1995) Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol.* 73, 218–226
36 DiCarlo, J.J. and Maunsell, J.H.R. (2003) Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *J. Neurophysiol.* 89, 3264–3278
37 Vogels, R. and Biederman, I. (2002) Effects of illumination intensity and direction on object coding in macaque inferior temporal cortex. *Cereb. Cortex* 12, 756–766
38 Zoccolan, D. *et al.* (2005) Multiple object response normalization in monkey inferotemporal cortex. *J. Neurosci.* 25, 8150–8164
39 DiCarlo, J.J. and Maunsell, J.H.R. (2000) Form representation in monkey inferotemporal cortex is virtually unaltered by free viewing. *Nat. Neurosci.* 3, 814–821
40 Moran, J. and Desimone, R. (1985) Selective attention gates visual processing in the extrastriate cortex. *Science* 229, 782–784
41 Suzuki, W. *et al.* (2006) Neuronal responses to object images in the macaque inferotemporal cortex at different stimulus discrimination levels. *J. Neurosci.* 26, 10524–10535
42 Naya, Y. *et al.* (2003) Delay-period activities in two subdivisions of monkey inferotemporal cortex during pair association memory task. *Eur. J. Neurosci.* 18, 2915–2918
43 McAdams, C.J. and Maunsell, J.H. (1999) Effects of attention on the reliability of individual neurons in monkey visual cortex. *Neuron* 23, 765–773
44 Reynolds, J.H. and Desimone, R. (2003) Interacting roles of attention and visual salience in V4. *Neuron* 37, 853–863
45 Chelazzi, L. *et al.* (1998) Responses of neurons in inferior temporal cortex during memory-guided visual search. *J. Neurophysiol.* 80, 2918–2940
46 Tsunoda, K. (2001) Complex objects are represented in macaque inferotemporal cortexby the combination of feature columns. *Nat. Neurosci.* 4, 832–838
47 Pasupathy, A. and Connor, C.E. (2001) Shape representation in area V4: position-specific tuning for boundary conformation. *J. Neurophysiol.* 86, 2505–2519
48 Brincat, S.L. and Connor, C.E. (2004) Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nat. Neurosci.* 7, 880–886
49 Yamane, Y. *et al.* (2006) Representation of the spatial relationship among object parts by neurons in macaque inferotemporal cortex. *J. Neurophysiol.* 96, 3147–3156
50 Kayaert, G. *et al.* (2003) Shape tuning in macaque inferior temporal cortex. *J. Neurosci.* 23, 3016–3027
51 Pollen, D.A. *et al.* (2002) Spatial receptive field organization of macaque V4 neurons. *Cereb. Cortex* 12, 601–616
52 Salinas, E. (2006) How behavioral constraints may determine optimal sensory representations. *PLoS Biol.* 4, e387
53 Cox, D.D. *et al.* (2005) 'Breaking' position-invariant object recognition. *Nat. Neurosci.* 8, 1145–1147
54 Wallis, G. and Bulthoff, H.H. (2001) Effects of temporal association on recognition memory. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4800–4804
55 Heeger, D.J. *et al.* (1996) Computational models of cortical visual processing. *Proc. Natl. Acad. Sci. U. S. A.* 93, 623–627
56 Einhauser, W. *et al.* (2005) Learning viewpoint invariant object representations using a temporal coherence principle. *Biol. Cybern.* 93, 79–90
57 Olshausen, B.A. and Field, D.J. (2004) Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* 14, 481–487
58 Olshausen, B.A. and Field, D.J. (2005) How close are we to understanding V1? *Neural Comput.* 17, 1665–1699

59 Simoncelli, E.P. (2003) Vision and the statistics of the visual environment. *Curr. Opin. Neurobiol.* 13, 144–149
60 Ullman, S. *et al.* (2002) Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.* 5, 682–687
61 Schwartz, O. and Simoncelli, E.P. (2001) Natural signal statistics and sensory gain control. *Nat. Neurosci.* 4, 819–825
62 Foldiak, P. (1991) Learning invariance from transformation sequences. *Neural Comput.* 3, 194–200
63 Wiskott, L. and Sejnowski, T.J. (2002) Slow feature analysis: unsupervised learning of invariances. *Neural Comput.* 14, 715–770
64 Edelman, S. and Intrator, N. (2003) Towards structural systematicity in distributed, statically bound visual representations. *Cognitive Sci.* 27, 73–110
65 Kersten, D. *et al.* (2004) Object perception as Bayesian inference. *Annu. Rev. Psychol.* 55, 271–304