Published in final edited form as:
Neuron. 2012 February 9; 73(3): 415–434. doi:10.1016/j.neuron.2012.01.010.

How does the brain solve visual object recognition?

James J. DiCarlo[1], **Davide Zoccolan**[2], and **Nicole C. Rust**[3]

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

[2]Cognitive Neuroscience and Neurobiology Sectors, International School for Advanced Studies (SISSA), Trieste, Italy

[3]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract

Mounting evidence suggests that "core object recognition," the ability to rapidly recognize objects despite substantial appearance variation, is solved in the brain via a cascade of reflexive, largely feedforward computations that culminate in a powerful neuronal representation in the inferior temporal cortex. However, the algorithm that produces this solution remains little-understood. Here we review evidence ranging from individual neurons, to neuronal populations, to behavior, to computational models. We propose that understanding this algorithm will require using neuronal and psychophysical data to sift through many computational models, each based on building blocks of small, canonical sub-networks with a common functional goal.

Introduction

Recognizing the words on this page, a coffee cup on your desk, or the person who just entered the room all seem so easy. The apparent ease of our visual recognition abilities belies the computational magnitude of this feat: we effortlessly detect and classify objects from among tens of thousands of possibilities (Biederman, 1987) and we do so within a fraction of a second (Potter, 1976; Thorpe et al., 1996), despite the tremendous variation in appearance that each object produces on our eyes (reviewed by Logothetis and Sheinberg, 1996). From an evolutionary perspective, our recognition abilities are not surprising -- our daily activities (e.g. finding food, social interaction, selecting tools, reading, etc.), and thus our survival, depends on our accurate and rapid extraction of object identity from the patterns of photons on our retinae.

The fact that half of the non-human primate neocortex is devoted to visual processing (Felleman and Van Essen, 1991) speaks to the computational complexity of object recognition. From this perspective, we have a remarkable opportunity -- we have access to a machine that produces a robust solution, and we can investigate that machine to uncover its algorithms of operation. These to-be-discovered algorithms will likely extend beyond the domain of vision -- not only to other biological senses (e.g. touch, audition, olfaction), but also to the discovery of meaning in high-dimensional artificial sensor data (e.g. cameras,

Correspondence to: James J. DiCarlo.

biometric sensors, etc.). Uncovering these algorithms requires expertise from psychophysics, cognitive neuroscience, neuroanatomy, neurophysiology, computational neuroscience, computer vision, and machine learning, and the traditional boundaries between these fields are dissolving.

What does it mean to say: "we want to understand object recognition"?

Conceptually, we want to know how the visual system can take each retinal image, and report the identities or categories of one or more objects that are present in that scene. Not everyone agrees on what a sufficient answer to object recognition might look like. One operational definition of "understanding" object recognition is the ability to construct an artificial system that performs as well as our own visual system (similar in spirit to computer-science tests of intelligence advocated by Turing (Turing, 1950). In practice, such an operational definition requires agreed-upon sets of images, tasks, and measures, and these "benchmark" decisions cannot be taken lightly (Pinto et al., 2008a; see below). The computer vision and machine learning communities might be content with a Turing definition of operational success, even if it looked nothing like the real brain, as it would capture useful computational algorithms independent of the hardware (or wetware) implementation. However, experimental neuroscientists tend to be more interested in mapping the spatial layout and connectivity of the relevant brain areas, uncovering conceptual definitions that can guide experiments, and reaching cellular and molecular targets that can be used to predictably modify object perception. For example, by uncovering the neuronal circuitry underlying object recognition, we might ultimately repair that circuitry in brain disorders that impact our perceptual systems (e.g. blindness, agnosias, etc.).

Nowadays, these motivations are synergistic -- experimental neuroscientists are providing new clues and constraints about the algorithmic solution at work in the brain, and computational neuroscientists seek to integrate these clues to produce hypotheses (a.k.a. algorithms) that can be experimentally distinguished. This synergy is leading to high-performing artificial vision systems (Pinto et al., 2008a; Pinto et al., 2009; Serre et al., 2007b). We expect this pace to accelerate, to fully explain human abilities, to reveal ways for extending and generalizing beyond those abilities, and to expose ways to repair broken neuronal circuits and augment normal circuits.

Progress toward understanding object recognition is driven by linking phenomena at different levels of abstraction. "Phenomena" at one level of abstraction (e.g., behavioral success on well-designed benchmark tests) are best explained by "mechanisms" at one level of abstraction below (e.g., a neuronal spiking population code in inferior temporal cortex, IT). Notably, these "mechanisms" are themselves "phenomena", that also require mechanistic explanations at an even lower level of abstraction (e.g., neuronal connectivity, intracellular events). Progress is facilitated by good intuitions about the most useful levels of abstraction as well as measurements of well-chosen phenomena at nearby levels. It then becomes crucial to define alternative hypotheses that link those sets of phenomena, and to determine those that explain the most data and generalize outside the specific conditions on which they were tested. In practice, we do not require all levels of abstraction and their links to be fully understood, but rather that both the phenomena and the linking hypotheses be understood sufficiently well as to achieve the broader policy missions of the research (e.g., building artificial vision systems, visual prosthetics, repairing disrupted brain circuits, etc.).

To that end, we review three sets of phenomena at three levels of abstraction (core recognition behavior, the IT population representation, and IT single unit responses), and we describe the links between these phenomena (Sections 1–2 below). We then consider how the architecture and plasticity of the ventral visual stream might produce a solution for

object recognition in IT (Sections 3), and we conclude by discussing key open directions (Section 4).

1. What is object recognition and why is it challenging?

The behavioral phenomenon of interest: core object recognition

Vision accomplishes many tasks besides object recognition, including object tracking, segmentation, obstacle avoidance, object grasping, etc., and these tasks are beyond the scope of this review. For example, studies point to the importance of the dorsal visual stream for supporting the ability to guide the eyes or covert processing resources (spatial "attention") toward objects (e.g. (Ikkai et al., 2011; Noudoost et al., 2010; Valyear et al., 2006) and to shape the hand to manipulate an object (e.g. (Goodale et al., 1994; Murata et al., 2000), and we do not review that work here (see (Cardoso-Leite and Gorea; Jeannerod et al., 1995; Konen and Kastner, 2008; Sakata et al., 1997). Instead, we and others define object recognition as the ability to assign labels (e.g., nouns) to particular objects, ranging from precise labels ("identification") to course labels ("categorization"). More specifically, we focus on the ability to complete such tasks over a range of identity preserving transformations (e.g., changes in object's position, size, pose, and background context), without any object-specific or location-specific pre-cuing (e.g., see Fig. 1). Indeed, primates can accurately report the identity or category of an object in the central visual field remarkably quickly: behavioral reaction times for single image presentations are as short as ~250 ms in monkeys (Fabre-Thorpe et al., 1998) and ~350 ms in humans (Rousselet et al., 2002; Thorpe et al., 1996), and images can be presented sequentially at rates less than ~100 ms per image (e.g. (Keysers et al., 2001; Potter, 1976). Accounting for the time needed to make a behavioral response, this suggests that the central visual image is processed to support recognition in less than 200 ms, even without attentional pre-cuing (Fabre-Thorpe et al., 1998; Intraub, 1980; Keysers et al., 2001; Potter, 1976; Rousselet et al., 2002; Rubin and Turano, 1992) Consistent with this, surface recordings in humans of evoked-potentials find neural signatures reflecting object identification within 150 ms (Thorpe et al., 1996). This "blink of an eye" time scale is not surprising in that primates typically explore their visual world with rapid eye movements, which result in short fixations (200–500 ms), during which the identity of one or more objects in the central visual field (~10 deg) must be rapidly determined. We refer to this extremely rapid and highly accurate object recognition behavior as "core recognition" (DiCarlo and Cox, 2007). This definition effectively strips the object recognition problem to its essence and provides a potentially tractable gateway to understanding. As describe below, it also places important constraints on the underlying neuronal codes (Section 2) and algorithms at work (Section 3).

The crux computational problem: core recognition requires invariance

To gain tractability, we have stripped the general problem of object recognition to the more specific problem of core recognition, but we have preserved its computational hallmark -- the ability to identify objects over a large range of viewing conditions. This so-called, "invariance problem" is *the* computational crux of recognition -- it is the major stumbling block for computer vision recognition systems (Pinto et al., 2008a; Ullman, 1996), particularly when many possible object labels must be entertained. The central importance of the invariance problem is easy to see if one imagines an engineer's task of building a recognition system for a visual world in which invariance was not needed. In such a world, repeated encounters of each object would evoke the same response pattern across the retina as previous encounters. In this world, object identity could easily be determined from the combined responses of the retinal population, and this procedure would easily scale to a nearly infinite number of possible "objects." This is not object recognition, and machine systems that work in these types of worlds already far outperform our own visual system.

In the real world, each encounter with an object is almost entirely unique, because of *identity-preserving image transformations*. Specifically, the vast array of images caused by objects that should receive the same label (e.g. “car”, Fig. 1) results from the variability of the world and the observer: each object can be encountered at any location on the retina (position variability), at a range of distances (scale variability), at many angles relative to the observer (pose variability), at a range lighting conditions (illumination variability), and in new visual contexts (clutter variability). Moreover, some objects are deformable in shape (e.g., bodies and faces), and often we need to group varying three-dimensional shapes into a common category such as “cars”, “faces” or “dogs” (intraclass variability). In sum, each encounter of the same object activates an entirely different retinal response pattern and the task of the visual system is to somehow establish the equivalence of all of these response patterns while, at the same time, not confuse any of them with images of all other possible objects (see Fig. 1).

Both behavioral (Potter, 1976; Thorpe et al., 1996) and neuronal (Hung et al., 2005) evidence suggest that the visual stream solves this invariance problem rapidly (discussed in Section 2). While the limits of such abilities have only been partly characterized (Afraz and Cavanagh, 2008; Bulthoff et al., 1995; Kingdom et al., 2007; Kravitz et al., 2010; Kravitz et al., 2008; Lawson, 1999; Logothetis et al., 1994b), from the point of view of an engineer, the brain achieves an impressive amount of invariance to identity-preserving image transformations (Pinto et al., 2010). Such invariance is a hallmark not only of primate vision, but is also found in evolutionary less advanced species (e.g., rodents; (Tafazoli et al., 2012; Zoccolan et al., 2009). In sum, the invariance of core object recognition is the right place to drive a wedge into the object recognition problem: it is operationally definable, it is a domain where biological visual systems excel, it is experimentally tractable, and it engages the crux computational difficulty of object recognition.

The invariance of core object recognition: a graphical intuition into the problem

A geometrical description of the invariance problem from a neuronal population coding perspective has been effective for motivating hypothetical solutions, including the notion that the ventral visual pathway gradually “untangles” information about object identity (DiCarlo and Cox, 2007). As a summary of those ideas, consider the response of a population of neurons to a particular view of one object as a response vector in a space whose dimensionality is defined by the number of neurons in the population (Fig. 2A). When an object undergoes an identity-preserving transformation, such as a shift in position or a change in pose, it produces a different pattern of population activity, which corresponds to a different response vector (Fig. 2A). Together, the response vectors corresponding to all possible identity-preserving transformations (e.g., changes in position, scale, pose, etc.) define a low-dimensional surface in this high dimensional space -- an object identity manifold (shown, for the sake of clarity, as a line in Fig. 2B). For neurons with small receptive fields that are activated by simple light patterns, such as retinal ganglion cells, each object manifold will be highly curved. Moreover, the manifolds corresponding to different objects will be “tangled” together, like pieces of paper crumpled into a ball (see Fig. 2B, left panel). At higher stages of visual processing, neurons tend to maintain their selectivity for objects across changes in view; this translates to manifolds that are more flat and separated (more “untangled”) (Fig 2B, right panel). Thus, object manifolds are thought to be gradually untangled through nonlinear selectivity and invariance computations applied at each stage of the ventral pathway (DiCarlo and Cox, 2007).

Object recognition is the ability to separate images that contain one particular object from images that do not (images of other possible objects; Fig. 1). In this geometrical perspective, this amounts to positioning a decision boundary, such as a hyperplane, to separate the manifold corresponding to one object from all other object manifolds. Mechanistically, one

can think of the decision boundary as approximating a higher order neuron that "looks down" on the population and computes object identity via a simple weighted sum of each neuron's responses, followed by a threshold. And thus it becomes clear why the representation at early stages of visual processing is problematic for object recognition: a hyperplane is completely insufficient for separating one manifold from the others because it is highly tangled with the other manifolds. However, at later stages, manifolds are flatter and not fused with each other, Fig. 2B), so that a simple hyperplane is all that is needed to separate them. This conceptual framework makes clear that information is not *created* as signals propagate through this visual system (which is impossible); rather, information is *reformatted* in a manner that makes information about object identity more explicit -- i.e., available to simple weighted summation decoding schemes. Later, we extend insights from object identity manifolds to how the ventral stream might accomplish this non-linear transformation.

Considering how the ventral stream might solve core recognition from this geometrical, population-based, perspective shifts emphasis away from traditional single neuron response properties, which display considerable heterogeneity in high-level visual areas and are difficult to understand (see Section 2). We argue that this perspective is a crucial intermediate level of understanding for the core recognition problem, akin to studying aerodynamics, rather than feathers, to understand flight. Importantly, this perspective suggests the immediate goal of determining how well each visual area has untangled the neuronal representation, which can be quantified via a simple summation decoding scheme (described above). It redirects emphasis toward determining the mechanisms that might contribute to untangling. And dictates what must be "explained" at the single neuron level, rather than creating "just so" stories based on the phenomenologies of heterogenous single neurons.

2. What do we know about the brain's "object" representation?

The ventral visual stream houses critical circuitry for core object recognition

Decades of evidence argue that the primate *ventral visual processing stream* -- a set of cortical areas arranged along the occipital and temporal lobes (Fig. 3A) -- houses key circuits that underlie object recognition behavior (For reviews, see Gross, 1994; Miyashita, 1993; Orban, 2008; Rolls, 2000). Object recognition is not the only ventral stream function, and we refer the reader to (Kravitz et al., 2010; Logothetis and Sheinberg, 1996; Maunsell and Treue, 2006; Tsao and Livingstone, 2008) for a broader discussion. Whereas lesions in the posterior ventral stream produce complete blindness in part of the visual field (reviewed by Stoerig and Cowey, 1997), lesions or inactivation of anterior regions, especially the inferior temporal cortex (IT), can produce selective deficits in the ability to distinguish among complex objects (e.g. (Holmes and Gross, 1984; Horel, 1996; Schiller, 1995; Weiskrantz and Saunders, 1984; Yaginuma et al., 1982). While these deficits are not always severe, and sometimes not found at all (Huxlin et al., 2000), this variability likely depends on the type of object recognition task (and thus the alternative visual strategies available). For example, some (Schiller, 1995; Weiskrantz and Saunders, 1984), but not all, primate ventral stream lesion studies have explicitly required invariance.

While the human homology to monkey IT cortex is not well-established, a likely homology is the cortex in and around the human lateral occipital cortex (LOC) (see (Orban et al., 2004) for review). For example, a comparison of monkey IT and human "IT" (LOC) shows strong commonality in the population representation of object categories (Kriegeskorte et al., 2008). Assuming these homologies, the importance of primate IT is suggested by neuropsychological studies of human patients with temporal lobe damage, which can sometimes produce remarkably specific object recognition deficits (Farah, 1990).

Temporary functional disruption of parts of the human ventral stream (using transcranial magnetic stimulation, TMS) can specifically disrupt certain types of object discrimination tasks, such as face discrimination (Pitcher et al., 2009). Similarly, artificial activation of monkey IT neurons predictably biases the subject's reported percept of complex objects (Afraz et al., 2006). In sum, long-term lesion studies, temporary activation/inactivation studies, and neurophysiological studies (described below) all point to the central role of the ventral visual stream in invariant object recognition.

Ventral visual stream: Multiple, hierarchically organized visual areas

The ventral visual stream has been parsed into distinct visual "areas" based on: anatomical connectivity patterns, distinctive anatomical structure, and retinotopic mapping (Felleman and Van Essen, 1991). Complete retinotopic maps have been revealed for most of the visual field (at least 40 degrees eccentricity from the fovea) for areas V1, V2 and V4 (Felleman and Van Essen, 1991) and thus each area can be thought of as conveying a population-based re-representation of each visually presented image. Within the IT complex, crude retinotopy exists over the more posterior portion (pIT; Boussaoud et al., 1991; Yasuda et al.), but retinotopy is not reported in the central and anterior regions (Felleman and Van Essen, 1991). Thus while IT is commonly parsed into sub-areas such as TEO and TE (Janssen et al., 2000; Saleem et al., 2000; Saleem et al., 1993; Suzuki et al., 2000; Von Bonin and P, 1947) or posterior IT (pIT), central IT (cIT) and anterior IT (aIT) (Felleman and Van Essen, 1991), it is unclear if IT cortex is more than one area, or how the term "area" should be applied. One striking illustration of this is recent monkey fMRI work which shows that there are three (Tsao et al., 2003) to six (Tsao et al., 2008a) or more (Ku et al., 2011) smaller regions within IT that may be involved in face "processing" (Tsao et al., 2008b) (also see (Op de Beeck et al., 2008; Pinsk et al., 2005). This suggests that, at the level of IT, behavioral goals (e.g. object categorization) (Kriegeskorte et al., 2008; Naselaris et al., 2009) many be a better spatial organizing principle than retinotopic maps.

All visual cortical areas share a six-layered structure and the inputs and outputs to each visual area share characteristic patterns of connectivity: ascending "feedforward" input is received in layer 4 and ascending "feedforward" output originates in the upper layers; descending "feedback" originates in the lower layers, and is received in the upper and lower layers of the "lower" cortical area (Felleman and Van Essen, 1991). These repeating connectivity patterns argue for a hierarchical organization (as opposed to a parallel or fully interconnected organization) of the areas with visual information traveling first from the retina to the lateral geniculate nucleus of the thalamus (LGN), and then through cortical area V1 to V2 to V4 to IT (Felleman and Van Essen, 1991). Consistent with this, the (mean) first visually evoked responses of each successive cortical area are successively lagged by ~10 ms (Nowak and Bullier, 1997; Schmolesky et al., 1998; see Fig. 3B). Thus just ~100 ms after image photons impinge on the retina, a first wave of image-selective neuronal activity is present throughout much of IT (e.g. (Desimone et al., 1984; DiCarlo and Maunsell, 2000; Hung et al., 2005; Kobatake and Tanaka, 1994a; Logothetis and Sheinberg, 1996; Tanaka, 1996). We believe this first wave of activity is consistent with a combination of intra-area processing and feed-forward inter-area processing of the visual image.

The ventral stream cortical code

The only known means of rapidly conveying information through the ventral pathway is via the spiking activity that travels along axons. Thus, we consider the neuronal representation in a given cortical area (e.g., the "IT representation") to be the spatiotemporal pattern of spikes produced by the set of pyramidal neurons that project out of that area (e.g. the spiking patterns traveling along the population of axons that project out of IT; see Fig. 3B). How is the spiking activity of individual neurons thought to encode visual information?

Most studies have investigated the response properties of neurons in the ventral pathway by assuming a firing rate (or, equivalently, a spike count) code, i.e., by counting how many spikes each neuron fires over several tens or hundreds of milliseconds following the presentation of a visual image, adjusted for latency (e.g., see Fig. 4A, B). Historically, this temporal window (here called the "decoding" window) was justified by the observation that its resulting spike rate is typically well modulated by relevant parameters of the presented visual images (such as object identity, position, or size; (Desimone et al., 1984; Kobatake and Tanaka, 1994b; Logothetis and Sheinberg, 1996; Tanaka, 1996) (see examples of IT neuronal responses in Fig. 4A–C), analogous to the well-understood firing rate modulation in area V1 by "low level" stimulus properties such as bar orientation (reviewed by Lennie and Movshon, 2005).

Like all cortical neurons, neuronal spiking throughout the ventral pathway is variable in the ms-scale timing of spikes, resulting in rate variability for repeated presentations of a nominally identical visual stimulus. This spike timing variability is consistent with a Poisson-like stochastic spike generation process with an underlying rate determined by each particular image (e.g. Kara et al., 2000; McAdams and Maunsell, 1999). Despite this variability, one can reliably infer what object, among a set of tested visual objects, was presented from the rates elicited across the IT population (e.g. (Abbott et al., 1996; Aggelopoulos and Rolls, 2005; De Baene et al., 2007; Heller et al., 1995; Hung et al., 2005; Li et al., 2006; Op de Beeck et al., 2001; Rust and DiCarlo, 2010). It remains unknown whether the ms-scale spike variability found in the ventral pathway is "noise" (in that it does not directly help stimulus encoding/decoding) or if it is somehow synchronized over populations of neurons to convey useful, perhaps "multiplexed" information (reviewed by Ermentrout et al., 2008). Empirically, taking into account the fine temporal structure of IT neuronal spiking patterns (e.g., concatenated decoding windows, each less than 50 ms) does not convey significantly more information about object identity than larger time windows (e.g. a single, 200 ms decoding window), suggesting that the results of ventral stream processing are well-described by a firing rate code where the relevant underlying time scale is ~50 ms (Abbott et al., 1996; Aggelopoulos and Rolls, 2005; Heller et al., 1995; Hung et al., 2005). While different time epochs relative to stimulus onset may encode different types of visual information (Brincat and Connor, 2006; Richmond and Optican, 1987; Sugase et al., 1999), very reliable object information is usually found in IT in the first ~50 ms of neuronal response (i.e. 100–150 ms after image onset, see Fig. 4A). More specifically, 1) the population representation is already different for different objects in that window (DiCarlo and Maunsell, 2000), and 2) that time window is more reliable because peak spike rates are typically higher than later windows (e.g. Hung et al., 2005). Deeper tests of ms-scale synchrony hypotheses require large-scale simultaneous recording. Another challenge to testing ms-scale spike coding is that alternative putative decoding schemes are typically unspecified and open-ended; a more complex scheme outside the range of each technical advance can always be postulated. In sum, while *all* spike-timing codes cannot easily (if ever) be ruled out, rate codes over ~50 ms intervals are not only easy to decode by downstream neurons, but appear to be sufficient to support recognition behavior (see below).

The IT population appears sufficient to support core object recognition

Although visual information processing in the first stage of the ventral stream (V1) is reasonably well understood (see (Lennie and Movshon, 2005) for review), processing in higher stages (e.g. V4, IT) remains poorly understood. Nevertheless, we know that the ventral stream produces an IT pattern of activity that can directly support robust, real-time visual object categorization and identification, even in the face of changes in object position and scale, limited clutter, and changes in background context (Hung et al., 2005; Li et al.,

2006; Rust and DiCarlo, 2010). Specifically, simple weighted summations of IT spike counts over short time intervals (see Section 2) lead to high rates of cross-validated performance for randomly selected populations of only a few hundred neurons (Hung et al., 2005; Rust and DiCarlo, 2010) (Fig. 4E), and a simple IT summation scheme is sufficient to explain a wide range of human invariant object recognition behavior (Majaj et al., 2012). Similarly, studies of fMRI-targeted clusters of IT neurons suggest that IT sub-populations can support other object recognition tasks such as face detection and face discrimination over some identity-preserving transformations (Freiwald and Tsao, 2010).

Importantly, IT neuronal populations are demonstrably better at object identification and categorization than populations at earlier stages of the ventral pathway (Freiwald and Tsao, 2010; Hung et al., 2005; Li et al., 2006; Rust and DiCarlo, 2010). Similarly, while neuronal activity that provides some discriminative information about object shape has also been found in dorsal stream visual areas at similar hierarchical levels (Sereno and Maunsell, 1998), a direct comparison shows that it is not nearly as powerful as IT for object discrimination (Lehky and Sereno, 2007).

Taken together, the neurophysiological evidence can be summarized as follows. First, spike counts in ~50 ms IT decoding windows convey information about visual object identity. Second, this information is available in the IT population beginning ~100 ms after image presentation (see Fig. 4A). Third, the IT neuronal representation of a given object across changes in position, scale, and presence of limited clutter is untangled from the representations of other objects, and object identity can be easily decoded using simple weighted summation codes (see Fig. 2B, 4D,E). Fourth, these codes are readily observed in passively viewing subjects, and for objects that have not been explicitly trained (Hung et al., 2005). In sum, our view is that the “output” of the ventral stream is reflexively expressed in neuronal firing rates across a short interval of time (~50 ms), is an “explicit” object representation (i.e., object identity is easily decodable), and the rapid production of this representation is consistent with a largely feedforward, non-linear processing of the visual input.

Alternative views suggest that ventral stream response properties are highly dependent on the subject’s behavioral state (i.e., “attention” or task goals) and that these state changes may be more appropriately reflected in global network properties (e.g., synchronized or oscillatory activity). While behavioral state effects, task effects, and plasticity have all been found in IT, such effects are typically (but not always) small relative to responses changes driven by changes in visual images (Koida and Komatsu, 2007; Op de Beeck and Baker, 2010; Suzuki et al., 2006; Vogels et al., 1995). Another, not-unrelated view is that true object representation is hidden in the fine-grained temporal spiking patterns of neurons and the correlational structure of those patterns. However, primate core recognition based on simple wighted summation of mean spike rates over 50–100 ms intervals is already powerful (Hung et al., 2005; Rust and DiCarlo, 2010), and appears to extend to difficult forms of invariance such as pose (Booth and Rolls, 1998; Freiwald and Tsao, 2010; Logothetis et al., 1995). More directly, decoded IT population performance exceeds artificial vision systems (Pinto et al., 2010; Serre et al., 2007a) and appears sufficient to explain human object recognition performance (Majaj et al., 2012). Thus, we work under the null hypothesis that core object recognition is well-described by a largely feedforward cascade of non-linear filtering operations (see below) and is expressed as a population rate code at ~50 ms time scale.

A contemporary view of IT single neurons

How do these IT neuronal population phenomena (above) depend on the responses of individual IT neurons? Understanding IT single-unit responses has proven to be extremely

challenging and while some progress has been made (Brincat and Connor, 2004; Yamane et al., 2008), we still have a poor ability to build encoding models that predict the responses of each IT neuron to new images (see Fig. 4B). Nevertheless, we know that IT neurons are activated by at least moderately complex combinations of visual features (Brincat and Connor, 2004; Desimone et al., 1984; Kobatake and Tanaka, 1994b; Perrett et al., 1982; Rust and DiCarlo, 2010; Tanaka, 1996), and that they are often able to maintain their relative object preference over small to moderate changes in object position and size (Brincat and Connor, 2004; Ito et al., 1995; Li et al., 2009; Rust and DiCarlo, 2010; Tovée et al., 1994), pose (Logothetis et al., 1994a), illumination (Vogels and Biederman, 2002) and clutter (Li et al., 2009; Missal et al., 1999; Missal et al., 1997; Zoccolan et al., 2005).

Contrary to popular depictions of IT neurons as narrowly selective “object detectors,” neurophysiological studies of IT are in near universal agreement with early accounts that describe a diversity of selectivity: “We found that, as in other visual areas, most IT neurons respond to many different visual stimuli and, thus, cannot be narrowly tuned “detectors” for particular complex objects...” (Desimone et al., 1984). For example, studies that involve probing the responses of IT cells with large and diverse stimulus sets show that, while some neurons appear highly selective for particular objects, they are the exception not the rule. Instead, most IT neurons are broadly tuned and the typical IT neuron responds to many different images and objects (Brincat and Connor, 2004; Freedman et al., 2006; Kreiman et al., 2006; Logothetis et al., 1995; Op de Beeck et al., 2001; Rolls, 1995, 2000; Rolls and Tovee, 1995; Vogels, 1999; Zoccolan et al., 2007); see Fig. 4B).

In fact, the IT population is diverse in both shape selectivity and tolerance to identity-preserving image transformations such as changes in object size, contrast, in-depth and in-plane rotation, and presence of background or clutter (Ito et al., 1995; Logothetis et al., 1995; Op de Beeck and Vogels, 2000; Perrett et al., 1982; Rust and DiCarlo, 2010; Zoccolan et al., 2005; Zoccolan et al., 2007). For example, the standard deviation of IT receptive field sizes is approximately 50% of the mean (mean ± SD: 16.5° ± 6.1° (Kobatake and Tanaka, 1994b), 24.5° ± 15.7° (Ito et al., 1995), and 10° ± 5° (Op de Beeck and Vogels, 2000). Moreover, IT neurons with the highest shape selectivities are the *least* tolerant to changes in position, scale, contrast and presence of visual clutter (Zoccolan et al., 2007), a finding inconsistent with “gnostic units” or “grandmother cells” (Gross, 2002), but one that arises naturally from feedforward computational models (Zoccolan et al., 2007).

Such findings argue for a distributed representation of visual objects in IT, as suggested previously (e.g. (Desimone et al., 1984; Kiani et al., 2007; Rolls, 1995) -- a view that motivates the population decoding approaches described above (Hung et al., 2005; Li et al., 2009; Rust and DiCarlo, 2010). That is, single IT neurons do not appear to act as sparsely active, invariant detectors of specific objects, but, rather, as elements of a population that, as a whole, supports object recognition. This implies that individual neurons do not need to be invariant. Instead, the key single-unit property is called neuronal “tolerance”: the ability of each IT neuron to maintain its preferences among objects, even if only over a limited transformation range (e.g., position changes; see Fig. 4C (Li et al., 2009). Mathematically, tolerance amounts to separable single-unit response surfaces for object shape and other object variables such as position and size (Brincat and Connor, 2004; Ito et al., 1995; Li et al., 2009; Tovée et al., 1994; see Fig. 4D). This contemporary view, that neuronal tolerance is the required and observed single unit phenomenology, has also been shown for less intuitive identity-preserving transformations such as the addition of clutter (Li et al., 2009; Zoccolan et al., 2005).

The tolerance of IT single units is non-trivial in that earlier visual neurons do not have this property to the same degree. It suggests that the IT neurons together tile the space of object

identity (shape) and other image variables such as object retinal position. The resulting population representation is powerful because it simultaneously conveys explicit information about object identity and its particular position, size, pose, and context, even when multiple objects are present, and it avoids the need to re-"bind" this information at a later stage (DiCarlo and Cox, 2007; Edelman, 1999; Riesenhuber and Poggio, 1999a). Graphically, this solution can be visualized as taking two sheets of paper (each is an object manifold) that are crumpled together, unfurling them, and aligning them on top of each other (DiCarlo and Cox, 2007). The surface coordinates of each sheet of paper correspond to identity-preserving object variables such as retinal position and, because they are aligned in this representation, this allows downstream circuits to use simple summation decoding schemes to answer questions such as: "Was there an object in the left visual field?", or "Which object was on the left?" (see Fig. 2B; DiCarlo and Cox, 2007).

3. What algorithm produces the IT population representation?

The results reviewed above argue that the ventral stream produces an IT population representation in which object identity and some other object variables (such as retinal position) are explicit, even in the face of significant image variation. But how is this achieved? Exactly what algorithm or set of algorithms is at work? We do not know the answer, but we have empirical data from neuroscience that partly constrains the hypothesis space, as well as computational frameworks that guide our intuition and show promise. In this section, we stand on those shoulders to speculate what the answer might look like.

The untangling solution is likely implemented in cortical circuitry

Retinal and LGN processing help deal with important real-world issues such as variation in luminance and contrast across each visual image (reviewed by Kohn, 2007). However, because RGC and LGN receptive fields are essentially point-wise spatial sensors (Field et al., 2010), the object manifolds conveyed to primary visual cortical area V1 are nearly as tangled as the pixel representation (see Fig. 2B). As V1 takes up the task, the number of output neurons, and hence the total dimensionality of the V1 representation, increases approximately thirty-fold (Stevens, 2001); Fig. 3B). Because V1 neuronal responses are non-linear with respect to their inputs (from the LGN), this dimensionality expansion results in an over-complete population re-representation (Lewicki and Sejnowski, 2000; Olshausen and Field, 1997) in which the object manifolds are more "spread out". Indeed, simulations show that a V1-like representation is clearly better than retinal-ganglion-cell-like (or pixel-based) representation, but still far below human performance for real-world recognition problems (DiCarlo and Cox, 2007; Pinto et al., 2008a).

Global scale architecture: a deep stack of cortical areas

What happens as each image is processed beyond V1 via the successive stages of the ventral stream anatomical hierarchy (V2, V4, pIT, aIT; Fig. 3)? Two overarching algorithmic frameworks have been proposed. One framework postulates that each successive visual area serially adds more processing power so as to solve increasingly complex tasks, such as the untangling of object identity manifolds (DiCarlo and Cox, 2007; Marr, 1982; Riesenhuber and Poggio, 1999b). A useful analogy here is a car assembly production line -- a single worker can only perform a small set of operations in a limited time, but a serial assembly line of workers can efficiently build something much more complex (e.g., a car or a good object representation).

A second algorithmic framework postulates the additional idea that the ventral stream hierarchy, and interactions between different levels of the hierarchy, embed important processing principles analogous to those in large hierarchical organizations, such as the US

Army (e.g. Lee and Mumford, 2003; Friston, 2010; Roelfsema and Houtkamp, 2011). In this framework, feedback connections between the different cortical areas are critical to the function of the system. This view has been advocated in part because it is one way to explicitly enable inference about objects in the image from weak or noisy data (e.g., missing or occluded edges) under a hierarchical Bayesian framework (Lee and Mumford, 2003; Rust and Stocker, 2010). For example, in the army analogy, foot soldiers (e.g. V1 neurons) pass uncertain observations (e.g. "maybe I see an edge") to sergeants (e.g. V2), who then pass the accumulated information to lieutenants, and so on. These higher agents thus glimpse the "forest for the trees" (e.g. Bar et al., 2006) and in turn direct the lowest levels (the foot soldiers) on how to optimize processing of this weak sensory evidence, presumably to help the higher agents (e.g. IT). A related, but distinct idea is that the hierarchy of areas plays a key role at a much slower time scale -- in particular, for *learning* to properly configure a largely-feedforward "serial chain" processing system (Hinton et al.,1995).

A central issue that separates the largely feedforward "serial-chain" framework and the feedforward/feedback "organized hierarchy" framework is whether reentrant areal communication (e.g. spikes sent from V1 to IT to V1) is necessary for building explicit object representation in IT within the time scale of natural vision (~200 ms). Even with improved experimental tools that might allow precise spatial-temporal shutdown of feedback circuits (e.g. Boyden et al., 2005), settling this debate hinges on clear predictions about the recognition tasks for which that reentrant processing is purportedly necessary. Indeed, it is likely that a compromise view is correct in that the best description of the system depends on the time scale of interest and the visual task conditions. For example, the visual system can be put in noisy or ambiguous conditions (e.g. binocular rivalry) in which coherent object percepts modulate on significantly slower time scales (seconds; e.g. Sheinberg and Logothetis, 1997) and this processing likely engages inter-area feedback along the ventral stream (e.g. Naya et al., 2001). Similarly, recognition tasks that involve extensive visual clutter (e.g. "Where's Waldo?") almost surely require overt reentrant processing (eye movements that cause new visual inputs) and/or covert feedback (Sheinberg and Logothetis, 2001; Ullman, 2009) as do working memory tasks that involve finding a specific object across a sequence of fixations (Engel and Wang, 2011). However, a potentially large class of object recognition tasks (what we call "core recognition", above) can be solved rapidly (~150 ms) and with the first spikes produced by IT (Hung et al., 2005; Thorpe et al., 1996), consistent with the possibility of little to no reentrant areal communication. Even if true, such data do not argue that core recognition is solved entirely by feedforward circuits -- very short time reentrant processing within spatially local circuits (<10 ms; e.g. local normalization circuits) is likely to be an integral part of the fast IT population response. Nor does it argue that anatomical pathways outside the ventral stream do not contribute to this IT solution (e.g. Bar et al., 2006). In sum, resolving debates about the necessity (or lack thereof) of reentrant processing in the areal hierarchy of ventral stream cortical areas depends strongly on developing agreed-upon operational definitions of "object recognition" (see Section 4), but the parsimonious hypothesis is that core recognition does not require reentrant areal processing.

Mesoscale architecture: inter-area and intra-area cortical relationships

One key idea implicit in both algorithmic frameworks is the idea of abstraction layers -- each level of the hierarchy need only be concerned with the "language" of its input area and its local job. For example, in the serial chain framework, while workers in the middle of a car assembly line might put in the car engine, they do not need to know the job description of early line workers (e.g., how to build a chassis). In this analogy, the mid-line workers are abstracted away from the job description of the early line workers.

Most complex, human-engineered systems have evolved to take advantage of abstraction layers, including the factory assembly line to produce cars and the reporting organization of large companies to produce coordinated action. Thus, the possibility that each cortical area can *abstract away* the details below its input area may be critical for leveraging a stack of visual areas (the ventral stream) to produce an untangled object identity representation (IT). A key advantage of such abstraction is that the "job description" of each worker is locally specified and maintained. The trade off is that, in its strongest instantiation, no one oversees the online operation of the entire processing chain and there are many workers at each level operating in parallel without explicit coordination (e.g., distant parts of V1). Thus, the proper upfront job description at each local cortical sub-population must be highly robust to that lack of across-area and within-area supervision. In principle, such robustness could arise from either an ultra precise, stable set of instructions given to each worker upfront (i.e., precise genetic control of *all* local cortical synaptic weights within the sub-population), or from a less precise "meta" job description -- initial instructions that are augmented by learning that continually refines the daily job description of each worker. Such learning mechanisms could involve feedback (e.g. Hinton et al., 1995; see above) and could act to refine the transfer function of each local sub-population.

Local architecture: Each cortical locus may have a common sub-space untangling goal

We argue above that the global function of the ventral stream might be best thought of as a collection of local input-output sub-populations (where each sub-population is a "worker") that are arranged laterally (to tile the visual field in each cortical area) and cascaded vertically (i.e. like an assembly line) with little or no need for coordination of those sub-populations at the time scale of online vision. We and others advocate the additional possibility that each ventral stream sub-population has an identical meta job description (see also Douglas and Martin, 1991; Fukushima, 1980, Kouh, 2008 #2493; Heeger et al., 1996). We say *"meta"* because we speculate about the implicit goal of each cortical sub-population, rather than its detailed transfer function (see below). This *canonical meta job description* would amount to an architectural scaffold and a set of learning rules describing how, following learning, the values of a finite number of inputs (afferents from lower cortical level) produce the values of a finite number of outputs (efferents to the next higher cortical level; see Fig. 5). We would expect these learning rules to operate at a much slower time scale than online vision. This possibility is not only conceptually simplifying to us as scientists, but it is extremely likely that an evolving system would exploit this type of computational unit because the same instruction set (e.g., genetic encoding of that meta job description) could simply be replicated laterally (to tile the sensory field) and stacked vertically (to gain necessary algorithmic complexity, see above). Indeed, while we have brought the reader here via arguments related to the processing power required for object representation, many have emphasized the remarkable architectural homogeneity of the mammalian neocortex (e.g., Douglas and Martin, 2004; Rockel et al., 1980); with some exceptions, each piece of neocortex copies many details of local structure (number of layers and cell types in each layer), internal connectivity (major connection statistics within that local circuit), and external connectivity (e.g., inputs from the lower cortical area arrive in layer 4, outputs to the next higher cortical area depart from layer 2/3).

For core object recognition, we speculate that the canonical meta job description of each local cortical sub-population is to solve a microcosm of the general untangling problem (Section 1). That is, instead of working on a ~1 million dimensional input basis, each cortical sub-population works on a much lower dimensional input basis (1–10 thousand; Fig. 5) which leads to significant advantages in both wiring packing and learnability from finite visual experience (Bengio, 2009). We call this hypothesized canonical meta goal *cortically local subspace untangling. "Cortically local" because* it is the hypothesized goal of every

local sub-population of neurons centered on any given point in ventral visual cortex (see Section 4), and "*subspace untangling*" because each such sub-population does not solve the full untangling problem, but instead aims to best untangle object identity within the data subspace afforded by its set of input afferents (e.g., a small aperture on the LGN in V1, a small aperture on V1 in V2, etc.). It is impossible for mossst cortical sub-populations to fully achieve this meta goal (because most only "see" a small window on each object), yet we believe that the combined efforts of many local units each trying their best to locally untangle may be all that is needed to produce an overall powerful ventral stream. That is, our hypothesis is that the parallel efforts of each ventral stream cortical locus to achieve local subspace untangling leads to a ventral stream assembly line whose "online" operation produces an untangled object representation at its top level. Later we outline how we aim to test that hypothesis.

"Bottom-up" encoding models of cortical responses

We have arrived at a putative canonical meta job description, local subspace untangling, by working our way "top down" from the overall goal of visual recognition and considering neuroanatomical data. How might local subspace untangling be instantiated within neuronal circuits and single neurons?

Historically, mechanistic insights into the computations performed by local cortical circuits have derived from "bottom up" approaches that aim to quantitatively describe the encoding functions that map image features to the firing rate responses of individual neurons. One example is the conceptual encoding models of Hubel and Wiesel (1962), which postulate the existence of two operations in V1 that produce the response properties of the "simple" and "complex" cells. First, V1 simple cells implement AND-like operations on LGN inputs to produce a new form of "selectivity" -- an orientation-tuned response. Next, V1 complex cells implement a form of "invariance" by making OR-like combinations of simple cells tuned for the same orientation. These conceptual models are central to current encoding models of biological object recognition (e.g. (Fukushima, 1980; Riesenhuber and Poggio, 1999b; Serre et al., 2007a), and they have been formalized into the linear-nonlinear (LN) class of encoding models in which each neuron adds and subtract its inputs, followed by a static nonlinearity (e.g., a threshold) to produce a firing rate response (Adelson and Bergen, 1985; Carandini et al., 2005; Heeger et al., 1996; Rosenblatt, 1958). While LN style models are far from a synaptic-level model of a cortical circuit, they are a potentially powerful level of abstraction in that they can account for a substantial amount of single neuron response patterns in early visual (Carandini et al., 2005), somatosensory (DiCarlo et al., 1998), and auditory cortical areas (Theunissen et al., 2000). Indeed, a nearly complete accounting of early level neuronal response patterns can be achieved with extensions to the simple LN model framework -- most notably, by divisive normalization schemes in which the output of each LN neuron is normalized (e.g. divided) by a weighted sum of a pool of nearby neurons (reviewed by Carandini and Heeger, 2011). Such schemes were used originally to capture luminance and contrast and other adaptation phenomena in the LGN and V1 (Mante et al., 2008; Rust and Movshon, 2005), and they represent a broad class of models which we refer to here as the "normalized LN" model class (NLN; see Fig. 5).

We do not know if the NLN class of encoding models can describe the local transfer function of any output neuron at any cortical locus (e.g., the transfer function from a V4 sub-population to a single IT neuron). However, because the NLN model is successful at the first sensory processing stage, the parsimonious view is to assume that the NLN model class is sufficient but that the particular NLN model parameters (i.e., the filter weights, the normalization pool, and the specific static non-linearity) of each neuron are uniquely elaborated. Indeed, the field has implicitly adopted this view with attempts to apply cascaded NLN-like models deeper into the ventral stream (e.g. David et al., 2006).

Unfortunately, the approach requires exponentially more stimulus-response data to try to constrain an exponentially expanding set of possible cascaded NLN models, and thus we cannot yet distinguish between a principled inadequacy of the cascaded NLN model class and a failure to obtain enough data. This is currently a severe "in practice" inadequacy of the cascaded NLN model class in that its effective explanatory power does not extend far beyond V1 (Carandini et al., 2005). Indeed, the problem of *directly* determining the specific *image-based* encoding function (e.g., a particular deep stack of NLN models) that predicts the response of any given IT neuron (e.g., the one at the end of my electrode today) may be practically impossible with current methods.

Canonical cortical algorithms: possible mechanisms of sub-space untangling

Nevertheless, all hope is not lost, and we argue for a different way forward. In particular, the appreciation of under-constrained models reminds us of the importance of abstraction layers in hierarchical systems -- returning to our earlier analogy, the workers at the end of the assembly line never need to build the entire car from scratch, but, together, the cascade of workers can still build a car. In other words, building an encoding model that describes the transformation from an image to a firing rate response is not the problem that (e.g.) an IT cortical neuron faces. On the contrary, the problem faced by each IT (NLN) neuron is a much more local, tractable, meta problem: from which V4 neurons should I receive inputs, how should I weigh them, what should comprise my normalization pool, and what static nonlinearity should I apply?.

Thus, rather than attempting to estimate the myriad parameters of each particular cascade of NLN models or each local NLN transfer function, we propose to focus instead on testing hypothetical meta job descriptions that can be implemented to produce those myriad details. We are particularly interested in hypotheses where the same (canonical) meta job description is invoked and set in motion at each cortical locus.

Our currently hypothesized meta job description (*cortically local subspace untangling)* is conceptually this: "Your job, as a local cortical sub-population, is to take all your neuronal afferents (your input representation) and apply a set of non-linearities and learning rules to adjust your input synaptic weights based on the activity of those afferents. These non-linearities and learning rules are designed such that, even though you do not know what an object is, your output representation will tend to be one in which object identity is more untangled than your input representation." Note that this is not a meta job description of each single neuron, but is the hypothesized goal of each local *sub-population* of neurons (see Fig. 5). It accepts that each neuron in the sub-population is well-approximated by a set of NLN parameters, but that many of these myriad parameters are highly idiosyncratic to each sub-population. Our hypothesis is that each ventral stream cortical sub-population uses at least three common, genetically encoded mechanisms (described below) to carry out that meta job description and that together, those mechanisms direct it to "choose" a set of input weights, a normalization pool, and a static nonlinearity that lead to improved subspace untangling. Specifically, we postulate the existence of the following three key conceptual mechanisms:

1. Each sub-population sets up architectural non-linearities that naturally tend to flatten object manifolds. Specifically, even with random (non-learned) filter weights, NLN-like models tend to produce easier-to-decode object identity manifolds largely on the strength of the normalization operation (Jarrett et al., 2009; Lewicki and Sejnowski, 2000; Olshausen and Field, 2005; Pinto et al., 2008b), similar in spirit to the overcomplete approach of V1 (described above).
2. Each sub-population embeds mechanisms that tune the synaptic weights to concentrate its dynamic response range to span regions of its input space where

images are typically found (e.g., do not bother encoding things you never see). This is the basis of natural image statistics and compression (e.g., Hoyer and Hyvarinen, 2002; Olshausen and Field, 1996; Simoncelli and Olshausen, 2001) and its importance is supported by the observation that higher levels of the ventral stream are more tuned to natural feature conjunctions than lower levels (e.g., Rust and DiCarlo, 2010).

3. Each sub-population uses an unsupervised algorithm to tune its parameters such that input patterns that occur close together in time tend to lead to similar output responses. This implements the theoretical idea that naturally occurring temporal contiguity cues can "instruct" the building of tolerance to identity-preserving transformations. More specifically, because each object's identity is temporally stable, different retinal images of the same object tend to be temporally contiguous (Fazl et al., 2009; Foldiak, 1991; Stryker, 1992; Wallis and Rolls, 1997; Wiskott and Sejnowski, 2002). In the geometrical, population-based description presented in Figure 2, response vectors that are produced by retinal images occurring close together in time tend to be the directions in the population response space that correspond to identity-preserving image variation, and thus attempts to produce similar neural responses for temporally contiguous stimuli achieve the larger goal of factorizing object identity and other object variables (position, scale, pose, etc.). For example, the ability of IT neurons to respond similarly to the same object seen at different retinal positions ("position tolerance") could be bootstrapped by the large number of saccadic-driven image translation experiences that are spontaneously produced on the retinae (~100 million such translation experiences per year of life). Indeed, artificial manipulations of temporally contiguous experience with object images across different positions and sizes can rapidly and strongly reshape the position and size tolerance of IT neurons -- destroying existing tolerance and building new tolerance, depending on the provided visual experience statistics (Li and DiCarlo, 2008, 2011), and predictably modifying object perception (Cox et al., 2005). We refer the reader to computational work on how such learning might explain properties of the ventral stream (e.g. Foldiak, 1991; Hurri and Hyvarinen, 2003; Wiskott and Sejnowski, 2002; see Section 4), as well as other potentially important types of unsupervised learning that do not require temporal cues (Karlinsky et al., 2008; Perry et al., 2010).

Testing hypotheses: instantiated models of the ventral stream

Experimental approaches are effective at describing undocumented behaviors of ventral stream neurons, but alone they cannot indicate when that search is complete. Similarly, "word models" (including ours, above) are not falsifiable algorithms. To make progress, we need to construct ventral-stream-inspired, instantiated computational models, and compare their performance with neuronal data and human performance on object recognition tasks. Thus, computational modeling cannot be taken lightly. Together, the set of alternative models define the space of falsifiable alternative hypotheses in the field, and the success of some such algorithms will be among our first indications that we are on the path to understanding visual object recognition in the brain.

The idea of using biologically inspired, hierarchical computational algorithms to understand the neuronal mechanisms underlying invariant object recognition tasks is not new: *"The mechanism of pattern recognition in the brain is little known, and it seems to be almost impossible to reveal it only by conventional physiological experiments... If we could make a neural network model which has the same capability for pattern recognition as a human being, it would give us a powerful clue to the understanding of the neural mechanism in the brain."* (Fukushima, 1980). More recent modeling efforts have significantly refined and

extended this approach (e.g. Lecun et al., 2004; Mel, 1997; Riesenhuber and Poggio, 1999b; Serre et al., 2007a). While we cannot review all the computer vision or neural network models that have relevance to object recognition in primates here, we refer the reader to reviews by (Bengio, 2009; Edelman, 1999; Riesenhuber and Poggio, 2000; Zhu and Mumford, 2006).

Commensurate with the serial chain, cascaded untangling discussion above, some ventral-stream-inspired models implement a canonical, iterated computation, with the overall goal of producing a good object representation at their highest stage (Fukushima, 1980; Riesenhuber and Poggio, 1999b; Serre et al., 2007a). These models include a handful of hierarchically arranged layers, each implementing AND-like operations to build selectivity followed by OR-like operations to build tolerance to identity preserving transformations (Fig. 6). Notably, both AND-like and OR-like computations can be formulated as variants of the NLN model class described above (Kouh and Poggio, 2008), illustrating the link to canonical cortical models (see inset in Fig. 6). Moreover, these relatively simple hierarchical models can produce model neurons that signal object identity, are somewhat tolerant to identity-preserving transformations, and can rival human performance for ultra-short, backward-masked image presentations (Serre et al., 2007a).

The surprising power of such models substantially demystifies the problem of invariant object recognition, but also points out that the devil is in the details -- the success of an algorithm depends on a large number of parameters that are only weakly constrained by existing neuroscience data. For example, while the algorithms of (Fukushima, 1980; Riesenhuber and Poggio, 1999b; Serre et al., 2007a) represent a great start, we also know that they are insufficient in that they perform only slightly better than baseline V1-like benchmark algorithms (Pinto et al., 2011), they fail to explain human performance for 100 ms or longer image presentations (Pinto et al., 2010), and their patterns of confusion do not match those found in the monkey IT representation (Kayaert et al., 2005; Kiani et al., 2007; Kriegeskorte et al., 2008). Nevertheless, these algorithms continue to inspire ongoing work, and recent efforts to more deeply explore the very large, ventral-stream-inspired algorithm class from which they are drawn is leading to even more powerful algorithms (Pinto et al., 2009), and motivating psychophysical testing and new neuronal data collection (Pinto et al., 2010; Majaj et al., 2012).

4. What is missing and how do we move forward?

Do we “understand” how the brain solves object recognition? We understand the computational crux of the problem (invariance); we understand the population coding issues resulting from invariance demands (object-identity manifold untangling); we understand where the brain solves this problem (ventral visual stream); and we understand the neuronal codes that are likely capable of supporting core recognition (~50 ms rate codes over populations of tolerant IT neurons). We also understand that the iteration of a basic class of largely feedforward functional units (NLN models configured as alternating patterns of AND-like and OR-like operations) can produce patterns of representations that approximate IT neuronal responses, produce respectable performance in computer vision tests of object recognition, and even approach some aspects of human performance. So what prevents us from declaring victory?

Problem 1. We must fortify intermediate levels of abstraction

At an elemental level, we have respectable models (e.g. NLN class; Heeger et al., 1996; Kouh and Poggio, 2008) of how each single unit computes its firing rate output from its inputs. However, we are missing a clear level of abstraction and linking hypotheses that can

connect mechanistic, NLN-like models to the resulting data reformatting that takes place in large neuronal populations (Fig. 5).

We argue that an iterative, canonical population processing motif provides a useful intermediate level of abstraction. The proposed canonical processing motif is intermediate in its physical instantiation (Fig. 5). Unlike NLN models, the canonical processing motif is a multi-input, multi-output circuit, with multiple afferents to layer 4 and multiple efferents from layer 2/3 and where the number of outputs is approximately the same as the number of inputs, thereby preserving the dimensionality of the local representation. We postulate the physical size of this motif to be ~500 um in diameter (~40K neurons), with ~10K input axons and ~10K output axons. This approximates the "cortical module" of Mountcastle (1997), and the "hypercolumn" of Hubel and Wiesel (1974), but is much larger than "ontogenetic microcolumns" suggested by neurodevelopment (Rakic, 1988) and the basic "canonical cortical circuit" (Douglas and Martin, 1991). The hypothesized sub-population of neurons is also intermediate in its algorithmic complexity. That is, unlike single NLN-like neurons, appropriately configured populations of (~10K) NLN-like neurons can, together, work on the type of population transformation that must be solved, but they cannot perform the task of the entire ventral stream. We propose that each processing motif has the same functional goal with respect to the patterns of activity arriving at its small input window; that is to use normalization architecture and unsupervised learning to factorize identity-preserving variables (e.g., position, scale, pose) from other variation (i.e., changes in object identity) in its input basis. As described above, we term this intermediate level processing motif "cortically local subspace untangling."

We must fortify this intermediate level of abstraction and determine if it provides the missing link. The next steps include: 1) We need to formally define "subspace untangling". Operationally, we mean that object identity will be easier to linearly decode on the output space than the input space, and we have some recent progress in that direction (Rust and DiCarlo, 2010). 2) We need to design and test algorithms that can qualitatively learn to produce the local untangling described in (1) and see if they also quantitatively produce the input-output performance of the ventral stream when arranged laterally (within an area) and vertically (across a stack of areas). There are a number of promising candidate ideas and algorithmic classes to consider (e.g. Hinton and Salakhutdinov, 2006; Olshausen and Field, 2004; Wiskott and Sejnowski, 2002). 3) We need to show how NLN-like models can be used to implement the learning algorithm in (2). In sum, we need to understand the relationship between intermediate-complexity algorithmic forms (e.g., filters with firing thresholds, normalization, competition, and unsupervised, time-driven associative learning) and manifold untangling (Fig. 2), as instantiated in local networks of ~40K cortical neurons.

Problem 2. The algorithmic solution lives in a very, very large space of "details"

We are not the first to propose a repeated cortical processing motif as an important intermediate abstraction. Indeed, some computational models adopt the notion of common processing motif, and make the same argument we reiterate here -- that an iterated application of a sub-algorithm is the correct way to think about the entire ventral stream (e.g., Fukushima, 1980; Kouh and Poggio, 2008; Riesenhuber and Poggio, 1999b; Serre et al., 2007a; see Fig. 6). However, no specific algorithm has yet achieved the performance of humans or explained the population behavior of IT (Pinto et al., 2011; Pinto et al., 2010).

The reason is that, while neuroscience has pointed to properties of the ventral stream that are likely critical to building explicit object representation (outlined above), there are many possible ways to instantiate such ideas as specific algorithms. For example, there are many possible ways to implement a series of AND-like operators followed by a series of OR-like operators, and it turns out that these details matter tremendously to the success or failure of

the resulting algorithm, both for recognition performance and for explaining neuronal data. Thus, these are not "details" of the problem -- understanding them *is* the problem.

Our proposal to solve this problem is to switch from inductive-style empirical science (where new neuronal data are used to motivate a new "word" model) to a systematic, quantitative search through the large class of possible algorithms, using experimental data to guide that search. In practice, we need to work in smaller algorithm spaces that use a reasonable number of meta-parameters to control a very large number of (e.g.) NLN-like parameters (see Section 3). For example, models that assume unsupervised learning use a small number of learning parameters to control a very large number of synaptic weight parameters (e.g. Bengio et al., 1995; Pinto et al., 2009; Serre et al., 2007b), which is one reason that neuronal evidence of unsupervised tolerance learning is of great interest to us (Section 3).

Exploration of these very large algorithmic classes is still in its infancy. However, we and our collaborators recently used rapidly advancing computing power to build many thousands of algorithms, in which a very large set of operating parameters was learned (unsupervised) from naturalistic video (Pinto et al., 2009). Optimized tests of object recognition (Pinto et al., 2008a) were then used to screen for the best algorithms. The resulting algorithms exceeded the performance of state-of-the-art computer vision models that had been carefully constructed over many years (Pinto et al., 2009). These very large, instantiated algorithm spaces are now being used to design large-scale neurophysiological recording experiments that aim to winnow out progressively more accurate models of the ventral visual stream.

Problem 3. We lack a systematic, operational definition of success

Although great strides have been made in biologically inspired vision algorithms (e.g. Hinton and Salakhutdinov, 2006; Lecun et al., 2004; Riesenhuber and Poggio, 1999b; Serre et al., 2007b; Ullman and Bart, 2004), the distance between human and computational algorithm performance remains poorly understood because there is little agreement on what the benchmarks should be. For example, one promising object recognition algorithm is competitive with humans under short presentations (20 ms) and backward-masked conditions, but its performance is still far below unfettered, 200 ms human core recognition performance (Serre et al., 2007a). How can we ask if an instantiated theory of primate object recognition is correct if we do not have an agreed upon definition of what "object recognition" is? Although we have given a loose definition (Section 1), a practical definition that can drive progress must operationally boil down to a strategy for generating sets of visual images or movies and defined tasks that can be measured in behavior, neuronal populations, and bio-inspired algorithms. This is easier said than done, as such tests must consider psychophysics, neuroscience, and computer vision; even supposed "natural, real-world" object recognition benchmarks do not easily distinguish between "state-of-the-art" computer vision algorithms and the algorithms that neuroscientists consider to be equivalent to a "null" model (e.g., performance of a crude model V1 population; Pinto et al.,2008b). Possible paths forward on the problem of benchmark tasks are outlined elsewhere (Pinto et al., 2008c), and the next steps require extensive psychophysical testing on those tasks to systematically characterize human abilities (e.g. Pinto et al., 2010; Majaj et al., 2012).

Problem 4. Synergies among the relevant domains of expertise must be nurtured

At a sociological level, progress has been challenged by the fact that the three most relevant research communities have historically been incentivized to focus on different objectives. Neuroscientists have focused on the problem of explaining the responses of individual neurons (e.g., Brincat and Connor, 2004; David et al., 2006) or mapping the locations of those neurons in the brain (e.g. Tsao et al., 2003), and using neuronal data to find algorithms

that explain human recognition performance has been only a hoped-for, but distant future outcome. For computer vision scientists that build object recognition algorithms, publication forces do not incentivize pointing out limitations or comparisons with older, simpler alternative algorithms. Moreover, the space of alternative algorithms is vague because industrial algorithms are not typically published, "new" object recognition algorithms from the academic community appear every few months, and there is little incentive to produce algorithms as downloadable, well documented code. Visual psychophysicists have traditionally worked in highly restricted stimulus domains and tasks that are thought to provide cleaner inference about the internal workings of the visual system. There is little incentive to systematically benchmark real-world object recognition performance for consumption by computational or experimental laboratories.

Fortunately, we are seeing increasing calls for meaningful collaboration by funding agencies, and collaborative groups are now working on all three pieces of the problem: 1) collecting the relevant psychophysical data, 2) collecting the relevant neuroscience data, and 3) putting together large numbers of alternative, instantiated computational models (algorithms) that work on real images (e.g., Cadieu et al., 2007; Zoccolan et al., 2007; Pinto et al., 2009; Pinto et al., 2010; Majaj et al., 2012).

Conclusion

We do not yet fully know how the brain solves object recognition. The first step is to clearly define the question itself. "Core object recognition," the ability to rapidly recognize objects in the central visual field in the face of image variation, is a problem that, if solved, will be the cornerstone for understanding biological object recognition. Although systematic characterizations of behavior are still ongoing, the brain has already revealed its likely solution to this problem in the spiking patterns of IT populations. Human-like levels of performance do not appear to require extensive recurrent communication, attention, task dependency or complex coding schemes that incorporate precise spike timing or synchrony. Instead, experimental and theoretical results remain consistent with this parsimonious hypothesis: a largely feedforward, reflexively computed, cascaded scheme in which visual information is gradually transformed and re-transmitted via a firing rate code along the ventral visual pathway, and presented for easy downstream consumption (i.e., simple weighted sums read out from the distributed population response).

To understand how the brain computes this solution, we must consider the problem at different levels of abstraction and the links between those levels. At the neuronal population level, the population activity patterns in early sensory structures that correspond to different objects are tangled together, but they are gradually untangled as information is re-represented along the ventral stream and in IT. At the single-unit level, this untangled IT object representation results from IT neurons that have some tolerance (rather than invariance) to identity-preserving transformations -- a property that neurons at earlier stages do not share, but that increases gradually along the ventral stream.

Understanding "how" the ventral pathway achieves this requires that we define one or more levels of abstraction between full cortical area populations and single neurons. For example, we hypothesize that canonical sub-networks of ~40K neurons form a basic "building block" for visual computation, and that each such sub-network has the same meta function. Even if this framework ultimately proves to be correct, it can only be shown by getting the many interacting "details" correct. Thus, progress will result from two synergistic lines of work. One line will use high-throughput computer simulations to systematically explore the very large space of possible sub-network algorithms, implementing each possibility as a cascaded, full scale algorithm, and measuring performance in carefully considered

benchmark object recognition tasks. A second line will used rapidly expanding systems neurophysiological data volumes and psychophysical performance measurements to sift through those algorithms for those that best explain the experimental data. Put simply, we must synergize the fields of psychophysics, systems neuroscience and computer vision around the problem of object recognition. Fortunately, the foundations and tools are now available to make it so.

Acknowledgments

JJD was supported by the US National Eye Institute (NIH NEI), The Defense Advanced Research Projects Agency (DARPA), and the National Science Foundation (NSF). DZ was supported by an Accademia Nazionale dei Lincei Compagnia di San Paolo Grant, a Programma Neuroscienze grant from the Compagnia di San Paolo, and a Marie Curie International Reintegration Grant. NR was supported by the US National Eye Institute (NIH NEI) and a fellowship from the Alfred P. Sloan Foundation.

References

Abbott LF, Rolls ET, Tovee MJ. Representational capacity of face coding in monkeys. Cerebral Cortex. 1996; 6:498–505. [PubMed: 8670675]

Adelson EH, Bergen JR. Spatiotemporal energy models for the perception of motion. J Opt Soc Am A. 1985; 2:284–299. [PubMed: 3973762]

Afraz SR, Cavanagh P. Retinotopy of the face aftereffect. Vision Res. 2008; 48:42–54. [PubMed: 18078975]

Afraz SR, Kiani R, Esteky H. Microstimulation of inferotemporal cortex influences face categorization. Nature. 2006; 442:692–695. [PubMed: 16878143]

Aggelopoulos NC, Rolls ET. Scene perception: inferior temporal cortex neurons encode the positions of different objects in the scene. Eur J Neurosci. 2005; 22:2903–2916. [PubMed: 16324125]

Bar M, Kassam KS, Ghuman AS, Boshyan J, Schmid AM, Dale AM, Hamalainen MS, Marinkovic K, Schacter DL, Rosen BR, Halgren E. Top-down facilitation of visual recognition. Proc Natl Acad Sci U S A. 2006; 103:449–454. [PubMed: 16407167]

Bengio Y. Learning deep architectures for AI. Foundations and Trends in Machine Learning. 2009; 2:1–127.

Bengio Y, LeCun Y, Nohl C, Burges C. LeRec: a NN/HMM hybrid for on-line handwriting recognition. Neural Comput. 1995; 7:1289–1303. [PubMed: 7584903]

Biederman I. Recognition-by-components: a theory of human image understanding. Psychol Rev. 1987; 94:115–147. [PubMed: 3575582]

Booth MCA, Rolls ET. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. Cerebral Cortex. 1998; 8:510–523. [PubMed: 9758214]

Boussaoud D, Desimone R, Ungerleider L. Visual topography of area TEO in the macaque. Journal of Comparative Neurology. 1991; 306:554–575. [PubMed: 1712794]

Boyden ES, Zhang F, Bamberg E, Nagel G, Deisseroth K. Millisecond-timescale, genetically targeted optical control of neural activity. Nat Neurosci. 2005; 8:1263–1268. [PubMed: 16116447]

Brewer AA, Press WA, Logothetis NK, Wandell BA. Visual areas in macaque cortex measured using functional magnetic resonance imaging. J Neurosci. 2002; 22:10416–10426. [PubMed: 12451141]

Brincat SL, Connor CE. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. Nat Neurosci. 2004; 7:880–886. [PubMed: 15235606]

Brincat SL, Connor CE. Dynamic shape synthesis in posterior inferotemporal cortex. Neuron. 2006; 49:17–24. [PubMed: 16387636]

Bulthoff HH, Edelman S, Tarr MJ. How are three-dimensional objects represented in the brain? Cerebral Cortex. 1995; 3:247–260. [PubMed: 7613080]

Cadieu C, Kouh M, Pasupathy A, Connor CE, Riesenhuber M, Poggio T. A model of V4 shape selectivity and invariance. J Neurophysiol. 2007; 98:1733–1750. [PubMed: 17596412]

Carandini M, Demb JB, Mante V, Tolhurst DJ, Dan Y, Olshausen BA, Gallant JL, Rust NC. Do we know what the early visual system does? J Neurosci. 2005; 25:10577–10597. [PubMed: 16291931]

Carandini M, Heeger DJ. Normalization as a canonical neural computation. Nat Rev Neurosci. 2011; 13:51–62. [PubMed: 22108672]

Cardoso-Leite P, Gorea A. On the perceptual/motor dissociation: a review of concepts, theory, experimental paradigms and data interpretations. Seeing Perceiving. 23:89–151. [PubMed: 20550823]

Collins CE, Airey DC, Young NA, Leitch DB, Kaas JH. Neuron densities vary across and within cortical areas in primates. Proc Natl Acad Sci U S A. 2010; 107:15927–15932. [PubMed: 20798050]

Cox DD, Meier P, Oertelt N, DiCarlo JJ. 'Breaking' position-invariant object recognition. Nat Neurosci. 2005; 8:1145–1147. [PubMed: 16116453]

David SV, Hayden BY, Gallant JL. Spectral receptive field properties explain shape selectivity in area V4. J Neurophysiol. 2006; 96:3492–3505. [PubMed: 16987926]

De Baene W, Premereur E, Vogels R. Properties of shape tuning of macaque inferior temporal neurons examined using rapid serial visual presentation. J Neurophysiol. 2007; 97:2900–2916. [PubMed: 17251368]

Desimone R, Albright TD, Gross CG, Bruce C. Stimulus-selective properties of inferior temporal neurons in the macaque. J Neurosci. 1984; 4:2051–2062. [PubMed: 6470767]

DiCarlo JJ, Cox DD. Untangling invariant object recognition. Trends Cogn Sci. 2007; 11:333–341. [PubMed: 17631409]

DiCarlo JJ, Johnson KO, Hsiao SS. Structure of receptive fields in area 3b of primary somatosensory cortex in the alert monkey. J Neurosci. 1998; 18:2626–2645. [PubMed: 9502821]

DiCarlo JJ, Maunsell JHR. Form representation in monkey inferotemporal cortex is virtually unaltered by free viewing. Nat Neurosci. 2000; 3:814–821. [PubMed: 10903575]

Douglas RJ, Martin KA. A functional microcircuit for cat visual cortex. J Physiol. 1991; 440:735–769. [PubMed: 1666655]

Douglas RJ, Martin KA. Neuronal circuits of the neocortex. Annu Rev Neurosci. 2004; 27:419–451. [PubMed: 15217339]

Edelman, S. Representation and Recognition in Vision. Cambridge, MA: MIT Press; 1999.

Engel TA, Wang XJ. Same or different? A neural circuit mechanism of similarity-based pattern match decision making. J Neurosci. 2011; 31:6982–6996. [PubMed: 21562260]

Ermentrout GB, Galan RF, Urban NN. Reliability, synchrony and noise. Trends Neurosci. 2008; 31:428–434. [PubMed: 18603311]

Fabre-Thorpe M, Richard G, Thorpe SJ. Rapid categorization of natural images by rhesus monkeys. Neuroreport. 1998; 9:303–308. [PubMed: 9507973]

Farah, MJ. Visual agnosia : disorders of object recognition and what they tell us about normal vision. Cambridge, Mass: MIT Press; 1990.

Fazl A, Grossberg S, Mingolla E. View-invariant object category learning, recognition, and search: how spatial and object attention are coordinated using surface-based attentional shrouds. Cogn Psychol. 2009; 58:1–48. [PubMed: 18653176]

Felleman DJ, Van Essen DC. Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex. 1991; 1:1–47. [PubMed: 1822724]

Field GD, Gauthier JL, Sher A, Greschner M, Machado TA, Jepson LH, Shlens J, Gunning DE, Mathieson K, Dabrowski W, et al. Functional connectivity in the retina at the resolution of photoreceptors. Nature. 2010; 467:673–677. [PubMed: 20930838]

Foldiak P. Learning invariance from transformation sequences. Neural Computation. 1991; 3:194–200.

Freedman DJ, Riesenhuber M, Poggio T, Miller EK. Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. Cereb Cortex. 2006; 16:1631–1644. [PubMed: 16400159]

Freiwald WA, Tsao DY. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. Science. 2010; 330:845–851. [PubMed: 21051642]

Friston K. The free-energy principle: a unified brain theory? Nat Rev Neurosci. 2010; 11:127–138. [PubMed: 20068583]

Fukushima K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern. 1980; 36:193–202. [PubMed: 7370364]

Goodale MA, Meenan JP, Bulthoff HH, Nicolle DA, Murphy KJ, Racicot CI. Separate neural pathways for the visual analysis of object shape in perception and prehension. Curr Biol. 1994; 4:604–610. [PubMed: 7953534]

Gross CG. How inferior temporal cortex became a visual area. Cereb Cortex. 1994; 4:455–469. [PubMed: 7833649]

Gross CG. Genealogy of the “grandmother cell”. Neuroscientist. 2002; 8:512–518. [PubMed: 12374433]

Heeger DJ, Simoncelli EP, Movshon JA. Computational models of cortical visual processing. Proc Natl Acad Sci U S A. 1996; 93:623–627. [PubMed: 8570605]

Heller J, Hertz JA, Kjaer TW, Richmond BJ. Information flow and temporal coding in primate pattern vision. J Computational Neuroscience. 1995; 2:175–193.

Hinton GE, Dayan P, Frey BJ, Neal RM. The “wake-sleep” algorithm for unsupervised neural networks. Science. 1995; 268:1158–1161. [PubMed: 7761831]

Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006; 313:504–507. [PubMed: 16873662]

Holmes EJ, Gross CG. Effects of inferior temporal lesions on discrimination of stimuli differing in orientation. J Neurosci. 1984; 4:3063–3068. [PubMed: 6502224]

Horel JA. Perception, learning and identification studied with reversible suppression of cortical visual areas in monkeys. Behav Brain Res. 1996; 76:199–214. [PubMed: 8734054]

Hoyer PO, Hyvarinen A. A multi-layer sparse coding network learns contour coding from natural images. Vision Res. 2002; 42:1593–1605. [PubMed: 12074953]

Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. Journal of Pysiology (London). 1962; 160:106–154.

Hubel DH, Wiesel TN. Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor. J Comp Neurol. 1974; 158:295–305. [PubMed: 4436457]

Hung CP, Kreiman G, Poggio T, DiCarlo JJ. Fast readout of object identity from macaque inferior temporal cortex. Science. 2005; 310:863–866. [PubMed: 16272124]

Hurri J, Hyvarinen A. Simple-cell-like receptive fields maximize temporal coherence in natural video. Neural Comput. 2003; 15:663–691. [PubMed: 12620162]

Huxlin KR, Saunders RC, Marchionini D, Pham HA, Merigan WH. Perceptual deficits after lesions of inferotemporal cortex in macaques [In Process Citation]. Cereb Cortex. 2000; 10:671–683. [PubMed: 10906314]

Ikkai A, Jerde TA, Curtis CE. Perception and action selection dissociate human ventral and dorsal cortex. J Cogn Neurosci. 2011; 23:1494–1506. [PubMed: 20465356]

Intraub H. Presentation rate and the representation of briefly glimpsed pictures in memory. J Exp Psychol [Hum Learn]. 1980; 6:1–12.

Ito M, Tamura H, Fujita I, Tanaka K. Size and position invariance of neuronal responses in monkey inferotemporal cortex. Journal of Neurophysiology. 1995; 73:218–226. [PubMed: 7714567]

Janssen P, Vogels R, Orban GA. Selectivity for 3D shape that reveals distinct areas within macaque inferior temporal cortex. Science. 2000; 288:2054–2056. [PubMed: 10856221]

Jarrett, K.; Kavukcuoglu, K.; Ranzato, M.; LeCun, Y. What is the Best Multi-Stage Architecture for Object Recognition?. Proc. International Conference on Computer Vision (ICCV’09); 2009.

Jeannerod M, Arbib MA, Rizzolatti G, Sakata H. Grasping objects: the cortical mechanisms of visuomotor transformation. Trends Neurosci. 1995; 18:314–320. [PubMed: 7571012]

Kara P, Reinagel P, Reid RC. Low response variability in simultaneously recorded retinal, thalamic, and cortical neurons. Neuron. 2000; 27:635–646. [PubMed: 11055444]

Karlinsky, L.; Michael, D.; Levi, D.; Ullman, S. Unsupervised classification and localization by consistency amplification. European Conference on Computer Vision; 2008. p. 321-335.

Kayaert G, Biederman I, Vogels R. Representation of regular and irregular shapes in macaque inferotemporal cortex. Cereb Cortex. 2005; 15:1308–1321. [PubMed: 15616128]

Keysers C, Xiao DK, Foldiak P, Perrett DI. The speed of sight. J Cogn Neurosci. 2001; 13:90–101. [PubMed: 11224911]

Kiani R, Esteky H, Mirpour K, Tanaka K. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. J Neurophysiol. 2007; 97:4296–4309. [PubMed: 17428910]

Kingdom FA, Field DJ, Olmos A. Does spatial invariance result from insensitivity to change? J Vis. 2007; 7(11):11–13. [PubMed: 18217806]

Kobatake E, Tanaka K. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J Neurophysiol. 1994a; 71:856–867. [PubMed: 8201425]

Kobatake E, Tanaka K. Neuronal selectivities to complex object-features in the ventral visual pathway of the macaque cerebral cortex. Journal of Neurophysiology. 1994b; 71:856–867. [PubMed: 8201425]

Kohn A. Visual adaptation: physiology, mechanisms, and functional benefits. J Neurophysiol. 2007; 97:3155–3164. [PubMed: 17344377]

Koida K, Komatsu H. Effects of task demands on the responses of color-selective neurons in the inferior temporal cortex. Nat Neurosci. 2007; 10:108–116. [PubMed: 17173044]

Konen CS, Kastner S. Two hierarchically organized neural systems for object information in human visual cortex. Nat Neurosci. 2008; 11:224–231. [PubMed: 18193041]

Kouh M, Poggio T. A canonical neural circuit for cortical nonlinear operations. Neural Comput. 2008; 20:1427–1451. [PubMed: 18254695]

Kravitz DJ, Kriegeskorte N, Baker CI. High-level visual object representations are constrained by position. Cereb Cortex. 2010; 20:2916–2925. [PubMed: 20351021]

Kravitz DJ, Vinson LD, Baker CI. How position dependent is visual object recognition? Trends Cogn Sci. 2008

Kreiman G, Hung CP, Kraskov A, Quiroga RQ, Poggio T, DiCarlo JJ. Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. Neuron. 2006; 49:433–445. [PubMed: 16446146]

Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron. 2008; 60:1126–1141. [PubMed: 19109916]

Ku SP, Tolias AS, Logothetis NK, Goense J. fMRI of the face-processing network in the ventral temporal lobe of awake and anesthetized macaques. Neuron. 2011; 70:352–362. [PubMed: 21521619]

Lawson R. Acheiving visual object constancy across plane rotation and depth rotation. Acta Psychol (Amst). 1999; 102:221–245. [PubMed: 10504882]

Lecun, Y.; Huang, F-J.; Bottou, L. Learning Methods for generic object recognition with invariance to pose and lighting. Proceedings of CVPR'04 (IEEE); 2004.

Lee TS, Mumford D. Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A Opt Image Sci Vis. 2003; 20:1434–1448. [PubMed: 12868647]

Lehky SR, Sereno AB. Comparison of shape encoding in primate dorsal and ventral visual pathways. J Neurophysiol. 2007; 97:307–319. [PubMed: 17021033]

Lennie P, Movshon JA. Coding of color and form in the geniculostriate visual pathway (invited review). J Opt Soc Am A Opt Image Sci Vis. 2005; 22:2013–2033. [PubMed: 16277273]

Lewicki MS, Sejnowski TJ. Learning overcomplete representations. Neural Comput. 2000; 12:337–365. [PubMed: 10636946]

Li, N.; Cox, DD.; Zoccolan, D.; DiCarlo, JJ. Flexible and robust object representation in inferior temporal cortex supported by neurons with limited position and clutter tolerance. Society for Neuroscience; Atlanta, GA: 2006.

Li N, Cox DD, Zoccolan D, DiCarlo JJ. What response properties do individual neurons need to underlie position and clutter "invariant" object recognition? J Neurophysiol. 2009

Li N, DiCarlo JJ. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science. 2008; 321:1502–1507. [PubMed: 18787171]

Li N, DiCarlo JJ. Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. Neuron. 2011; 67:1062–1075. [PubMed: 20869601]

Logothetis NK, Pauls J, Bulthoff HH, Poggio T. View-dependent object recognition by monkeys. Curr Biol. 1994a; 4:401–414. [PubMed: 7922354]

Logothetis, NK.; Pauls, J.; Poggio, T. Viewer-centered object recognition in monkeys. MIT; 1994b.

Logothetis NK, Pauls J, Poggio T. Shape representation in the inferior temporal cortex of monkeys. Curr Biol. 1995; 5:552–563. [PubMed: 7583105]

Logothetis NK, Sheinberg DL. Visual object recognition. Ann Rev Neurosci. 1996; 19:577–621. [PubMed: 8833455]

Majaj, N.; Najib, H.; Solomon, E.; DiCarlo, JJ. A unified neuronal population code fully explains human object recognition. Computational and Systems Neuroscience (COSYNE); Salt Lake City, UT: 2012.

Mante V, Bonin V, Carandini M. Functional mechanisms shaping lateral geniculate responses to artificial and natural stimuli. Neuron. 2008; 58:625–638. [PubMed: 18498742]

Marr, D. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Henry Holt & Company; 1982.

Maunsell JH, Treue S. Feature-based attention in visual cortex. Trends Neurosci. 2006; 29:317–322. [PubMed: 16697058]

McAdams CJ, Maunsell JH. Effects of attention on the reliability of individual neurons in monkey visual cortex [In Process Citation]. Neuron. 1999; 23:765–773. [PubMed: 10482242]

Mel BW. SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. Neural Comput. 1997; 9:777–804. [PubMed: 9161022]

Missal M, Vogels R, Li C, Orban GA. Shape interactions in macaque inferior temporal neurons. Journal of Neurophysiology. 1999; 82:131–142. [PubMed: 10400942]

Missal M, Vogels R, Orban GA. Responses of macaque inferior temporal neurons to overlapping shapes. Cereb Cortex. 1997; 7:758–767. [PubMed: 9408040]

Miyashita Y. Inferior temporal cortex: where visual perception meets memory. Annual Review of Neuroscience. 1993; 16:245–263.

Mountcastle VB. The columnar organization of the neocortex. Brain. 1997; 120(Pt 4):701–722. [PubMed: 9153131]

Murata A, Gallese V, Luppino G, Kaseda M, Sakata H. Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area AIP. J Neurophysiol. 2000; 83:2580–2601. [PubMed: 10805659]

Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL. Bayesian reconstruction of natural images from human brain activity. Neuron. 2009; 63:902–915. [PubMed: 19778517]

Naya Y, Yoshida M, Miyashita Y. Backward Spreading of Memory-Retrieval Signal in the Primate Temporal Cortex. Science. 2001; 291:661–664. [PubMed: 11158679]

Noudoost B, Chang MH, Steinmetz NA, Moore T. Top-down control of visual attention. Curr Opin Neurobiol. 2010; 20:183–190. [PubMed: 20303256]

Nowak, LG.; Bullier, J. The timing of information transfer in the visual system. In: Rockland, K.; Kaas, J.; Peters, A., editors. Cerebral Cortex: Extrastriate Cortex in Primate. Plenum Publishing Corporation; 1997. p. 870

O'Kusky J, Colonnier M. A laminar analysis of the number of neurons, glia, and synapses in the adult cortex (area 17) of adult macaque monkeys. J Comp Neurol. 1982; 210:278–290. [PubMed: 7142443]

Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images [see comments]. Nature. 1996; 381:607–609. [PubMed: 8637596]

Olshausen BA, Field DJ. Sparse coding with an overcomplete basis set: a strategy employed by V1? Vision Res. 1997; 37:3311–3325. [PubMed: 9425546]

Olshausen BA, Field DJ. Sparse coding of sensory inputs. Curr Opin Neurobiol. 2004; 14:481–487. [PubMed: 15321069]

Olshausen BA, Field DJ. How close are we to understanding v1? Neural Comput. 2005; 17:1665–1699. [PubMed: 15969914]

Op de Beeck H, Vogels R. Spatial sensitivity of macaque inferior temporal neurons. J Comp Neurol. 2000; 426:505–518. [PubMed: 11027395]

Op de Beeck H, Wagemans J, Vogels R. Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. Nat Neurosci. 2001; 4:1244–1252. [PubMed: 11713468]

Op de Beeck HP, Baker CI. Informativeness and learning: Response to Gauthier and colleagues. Trends Cogn Sci. 2010; 14:236–237. [PubMed: 20714344]

Op de Beeck HP, DiCarlo JJ, Goense JB, Grill-Spector K, Papanastassiou A, Tanifuji M, Tsao DY. Fine-scale spatial organization of face and object selectivity in the temporal lobe: do functional magnetic resonance imaging, optical imaging, and electrophysiology agree? J Neurosci. 2008; 28:11796–11801. [PubMed: 19005042]

Orban GA. Higher order visual processing in macaque extrastriate cortex. Physiol Rev. 2008; 88:59–89. [PubMed: 18195083]

Orban GA, Van Essen D, Vanduffel W. Comparative mapping of higher visual areas in monkeys and humans. Trends Cogn Sci. 2004; 8:315–324. [PubMed: 15242691]

Perrett DI, Rolls ET, Caan W. Visual neurones responsive to faces in the monkey temporal cortex. Experimental Brain Research. 1982; 47:329–342.

Perry G, Rolls ET, Stringer SM. Continuous transformation learning of translation invariant representations. Exp Brain Res. 2010; 204:255–270. [PubMed: 20544186]

Pinsk MA, DeSimone K, Moore T, Gross CG, Kastner S. Representations of faces and body parts in macaque temporal cortex: a functional MRI study. Proc Natl Acad Sci U S A. 2005; 102:6996–7001. [PubMed: 15860578]

Pinto, N.; Barhomi, Y.; Cox, DD.; DiCarlo, JJ. Comparing State-of-the-Art Visual Features on Invariant Object Recognition Tasks. IEEE Workshop on Applications of Computer Vision; Kona, HI. 2011.

Pinto, N.; Cox, DD.; Corda, B.; Doukhan, D.; DiCarlo, JJ. Why is real-world object recognition hard?: Establishing honest benchmarks and baselines for object recognition. COSYNE; Salt Lake City, UT: 2008a.

Pinto N, Cox DD, DiCarlo JJ. Why is real-world visual object recognition hard? PLoS Comput Biol. 2008b; 4:e27. [PubMed: 18225950]

Pinto, N.; DiCarlo, J.; Cox, D. Establishing Benchmarks and Baselines for Face Recognition. ECCV 2008 Faces in Real Life Workshop; Marseille, France. 2008c.

Pinto N, Doukhan D, DiCarlo JJ, Cox DD. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. PLoS Comput Biol. 2009; 5:e1000579. [PubMed: 19956750]

Pinto, N.; Majaj, N.; YB; EAS; DDC; DiCarlo, J. Human versus machine: comparing visual object recognition systems on a level playing field. F.N.C.A.C.a.S.N; 2010.

Pitcher D, Charles L, Devlin JT, Walsh V, Duchaine B. Triple dissociation of faces, bodies, and objects in extrastriate cortex. Curr Biol. 2009; 19:319–324. [PubMed: 19200723]

Potter MC. Short-term conceptual memory for pictures. J Exp Psychol [Hum Learn]. 1976; 2:509–522.

Rakic P. Specification of cerebral cortical areas. Science. 1988; 241:170–176. [PubMed: 3291116]

Richmond BJ, Optican LM. Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. II. Quantification of response waveform. Journal of Neurophysiology. 1987; 57:147–161. [PubMed: 3559669]

Riesenhuber M, Poggio T. Are cortical models really bound by the “binding problem”? Neuron. 1999a; 24:87–93. 111–125. [PubMed: 10677029]

Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. Nat Neurosci. 1999b; 2:1019–1025. [PubMed: 10526343]

Riesenhuber M, Poggio T. Models of object recognition. Nat Neurosci. 2000; 3(Suppl):1199–1204. [PubMed: 11127838]

Rockel AJ, Hiorns RW, Powell TP. The basic uniformity in structure of the neocortex. Brain. 1980; 103:221–244. [PubMed: 6772266]

Roelfsema PR, Houtkamp R. Incremental grouping of image elements in vision. Atten Percept Psychophys. 2011; 73:2542–2572. [PubMed: 21901573]

Rolls ET. Sparseness of the Neuronal Representation of Stimuli in the Primate Temporal Visual Cortex. Journal of Neurophysiology. 1995; 73:713–726. [PubMed: 7760130]

Rolls ET. Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. Neuron. 2000; 27:205–218. [PubMed: 10985342]

Rolls ET, Tovee MJ. Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. Journal of Neurophysiology. 1995; 73:713–726. [PubMed: 7760130]

Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958; 65:386–408. [PubMed: 13602029]

Rousselet GA, Fabre-Thorpe M, Thorpe SJ. Parallel processing in high-level categorization of natural images. Nat Neurosci. 2002; 5:629–630. [PubMed: 12032544]

Rubin GS, Turano K. Reading without saccadic eye movements. Vision Res. 1992; 32:895–902. [PubMed: 1604858]

Rust NC, DiCarlo JJ. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. J Neurosci. 2010; 30:12978–12995. [PubMed: 20881116]

Rust NC, Movshon JA. In praise of artifice. Nat Neurosci. 2005; 8:1647–1650. [PubMed: 16306892]

Rust NC, Stocker AA. Ambiguity and invariance: two fundamental challenges for visual processing. Curr Opin Neurobiol. 2010

Sakata H, Taira M, Kusunoki M, Murata A, Tanaka Y. The TINS Lecture. The parietal association cortex in depth perception and visual control of hand action. Trends Neurosci. 1997; 20:350–357. [PubMed: 9246729]

Saleem KS, Suzuki W, Tanaka K, Hashikawa T. Connections between anterior inferotemporal cortex and superior temporal sulcus regions in the macaque monkey [In Process Citation]. J Neurosci. 2000; 20:5083–5101. [PubMed: 10864966]

Saleem KS, Tanaka K, Rockland KS. Specific and columnar projection from area TEO to TE in the macaque inferotemporal cortex. Cereb Cortex. 1993; 3:454–464. [PubMed: 8260813]

Schiller PH. Effect of lesion in visual cortical area V4 on the recognition of transformed objects. Nature. 1995; 376:342–344. [PubMed: 7630401]

Schmolesky MT, Wang Y, Hanes DP, Thompson KG, Leutgeb S, Schall JD, Leventhal AG. Signal timing across the macaque visual system. J Neurophysiol. 1998; 79:3272–3278. [PubMed: 9636126]

Sereno AB, Maunsell JH. Shape selectivity in primate lateral intraparietal cortex [see comments]. Nature. 1998; 395:500–503. [PubMed: 9774105]

Serre T, Oliva A, Poggio T. A feedforward architecture accounts for rapid categorization. Proc Natl Acad Sci U S A. 2007a; 104:6424–6429. [PubMed: 17404214]

Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T. Robust object recognition with cortex-like mechanisms. IEEE Trans Pattern Anal Mach Intell. 2007b; 29:411–426. [PubMed: 17224612]

Sheinberg DL, Logothetis NK. The role of temporal cortical areas in perceptual organization. Proc Natl Acad Sci U S A. 1997; 94:3408–3413. [PubMed: 9096407]

Sheinberg DL, Logothetis NK. Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. J Neurosci. 2001; 21:1340–1350. [PubMed: 11160405]

Simoncelli EP, Olshausen BA. Natural image statistics and neural representation. Annu Rev Neurosci. 2001; 24:1193–1216. [PubMed: 11520932]

Stevens CF. An evolutionary scaling law for the primate visual system and its basis in cortical function. Nature. 2001; 411:193–195. [PubMed: 11346795]

Stoerig P, Cowey A. Blindsight in man and monkey. Brain. 1997; 120(Pt 3):535–559. [PubMed: 9126063]

Stryker MP. Neurobiology. Elements of visual perception. Nature. 1992; 360:301–302. [PubMed: 1448145]

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. [PubMed: 10476965]

Suzuki W, Matsumoto K, Tanaka K. Neuronal responses to object images in the macaque inferotemporal cortex at different stimulus discrimination levels. J Neurosci. 2006; 26:10524–10535. [PubMed: 17035537]

Suzuki W, Saleem KS, Tanaka K. Divergent backward projections from the anterior part of the inferotemporal cortex (area TE) in the macaque. J Comp Neurol. 2000; 422:206–228. [PubMed: 10842228]

Tafazoli S, Di Filippo A, Zoccolan D. Transformation-Tolerant Object Recognition in Rats Revealed by Visual Priming. J Neurosci. 2012; 32(1):21–34. [PubMed: 22219267]

Tanaka K. Inferotemporal cortex and object vision. Annual Review of Neuroscience. 1996; 19:109–139.

Theunissen FE, Sen K, Doupe AJ. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. J Neurosci. 2000; 20:2315–2331. [PubMed: 10704507]

Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system. Nature. 1996; 381:520–522. [PubMed: 8632824]

Tovée MJ, Rolls ET, Azzopardi P. Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert monkey. Journal of Neurophysiology. 1994; 72:1049–1060. [PubMed: 7807195]

Tsao DY, Freiwald WA, Knutsen TA, Mandeville JB, Tootell RB. Faces and objects in macaque cerebral cortex. Nat Neurosci. 2003; 6:989–995. [PubMed: 12925854]

Tsao DY, Livingstone MS. Mechanisms of face perception. Annu Rev Neurosci. 2008; 31:411–437. [PubMed: 18558862]

Tsao DY, Moeller S, Freiwald WA. Comparing face patch systems in macaques and humans. Proc Natl Acad Sci U S A. 2008a; 105:19514–19519. [PubMed: 19033466]

Tsao DY, Schweers N, Moeller S, Freiwald WA. Patches of face-selective cortex in the macaque frontal lobe. Nat Neurosci. 2008b; 11:877–879. [PubMed: 18622399]

Turing AM. Computing Machinery and Intelligence. 49. Mind. 1950; 49:433–460.

Ullman, S. High Level Vision. Cambridge, MA: MIT Press; 1996.

Ullman, S. Beyond Classification. In: Dickinson, editor. Object Categorization: Compuer and Human Vision Perspectives. Cambridge University Press; 2009.

Ullman S, Bart E. Recognition invariance obtained by extended and invariant features. Neural Netw. 2004; 17:833–848. [PubMed: 15288901]

Valyear KF, Culham JC, Sharif N, Westwood D, Goodale MA. A double dissociation between sensitivity to changes in object identity and object orientation in the ventral and dorsal visual streams: a human fMRI study. Neuropsychologia. 2006; 44:218–228. [PubMed: 15955539]

Vogels R. Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. Eur J Neurosci. 1999; 11:1239–1255. [PubMed: 10103119]

Vogels R, Biederman I. Effects of illumination intensity and direction on object coding in macaque inferior temporal cortex. Cereb Cortex. 2002; 12:756–766. [PubMed: 12050087]

Vogels R, Sáry G, Orban GA. How task-related are the responses of inferior temporal neurons? Visual Neuroscience. 1995; 12:207–214. [PubMed: 7786842]

Von Bonin, G.; PB. The neocortex of Macaca mulatta. Urbana, IL: University of Illinois Press; 1947.

Wallis G, Rolls ET. Invariant face and object recognition in the visual system. Progress in Neurobiology. 1997; 51:167–194. [PubMed: 9247963]

Weiskrantz L, Saunders RC. Impairments of visual object transforms in monkeys. Brain. 1984; 107:1033–1072. [PubMed: 6509307]

Wiskott L, Sejnowski TJ. Slow feature analysis: unsupervised learning of invariances. Neural Comput. 2002; 14:715–770. [PubMed: 11936959]

Yaginuma S, Niihara T, Iwai E. Further evidence on elevated discrimination limens for reduced patterns in monkeys with inferotemporal lesions. Neuropsychologia. 1982; 20:21–32. [PubMed: 7070649]

Yamane Y, Carlson ET, Bowman KC, Wang Z, Connor CE. A neural code for three-dimensional object shape in macaque inferotemporal cortex. Nat Neurosci. 2008

Yasuda M, Banno T, Komatsu H. Color selectivity of neurons in the posterior inferior temporal cortex of the macaque monkey. Cereb Cortex. 20:1630–1646. [PubMed: 19880593]

Zhu S, Mumford D. A stochastic grammar of images. Foundations and Trends in Computer Graphics and Vision. 2006; 2:259–362.

Zoccolan D, Cox DD, DiCarlo JJ. Multiple object response normalization in monkey inferotemporal cortex. J Neurosci. 2005; 25:8150–8164. [PubMed: 16148223]

Zoccolan D, Kouh M, Poggio T, DiCarlo JJ. Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. J Neurosci. 2007; 27:12292–12307. [PubMed: 17989294]

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. X. [PubMed: 19429704]

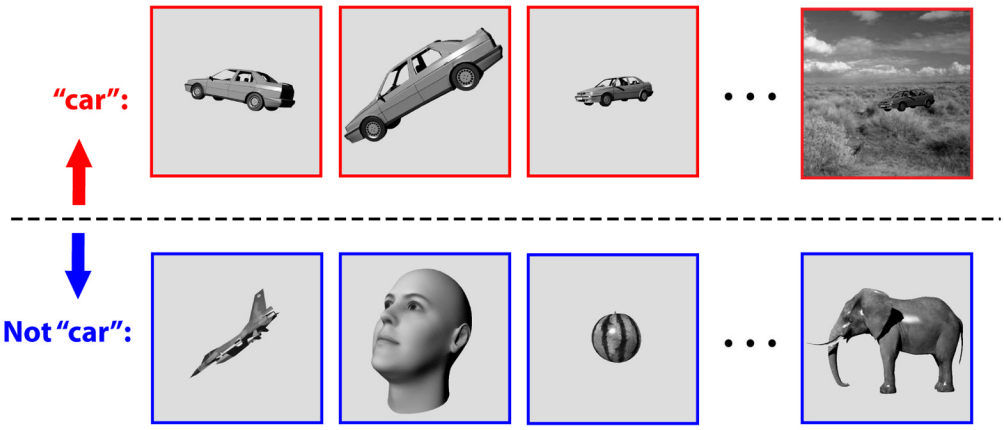

Figure 1. Core object recognition

is the ability to rapidly (<200 ms viewing duration) discriminate a given visual object (e.g., a car, top row) from all other possible visual objects (e.g. bottom row) without any object-specific or location-specific pre-cuing (e.g. (DiCarlo and Cox, 2007). Primates perform this task remarkably well, even in the face of identity-preserving transformations (e.g., changes in object position, size, viewpoint, and visual context).

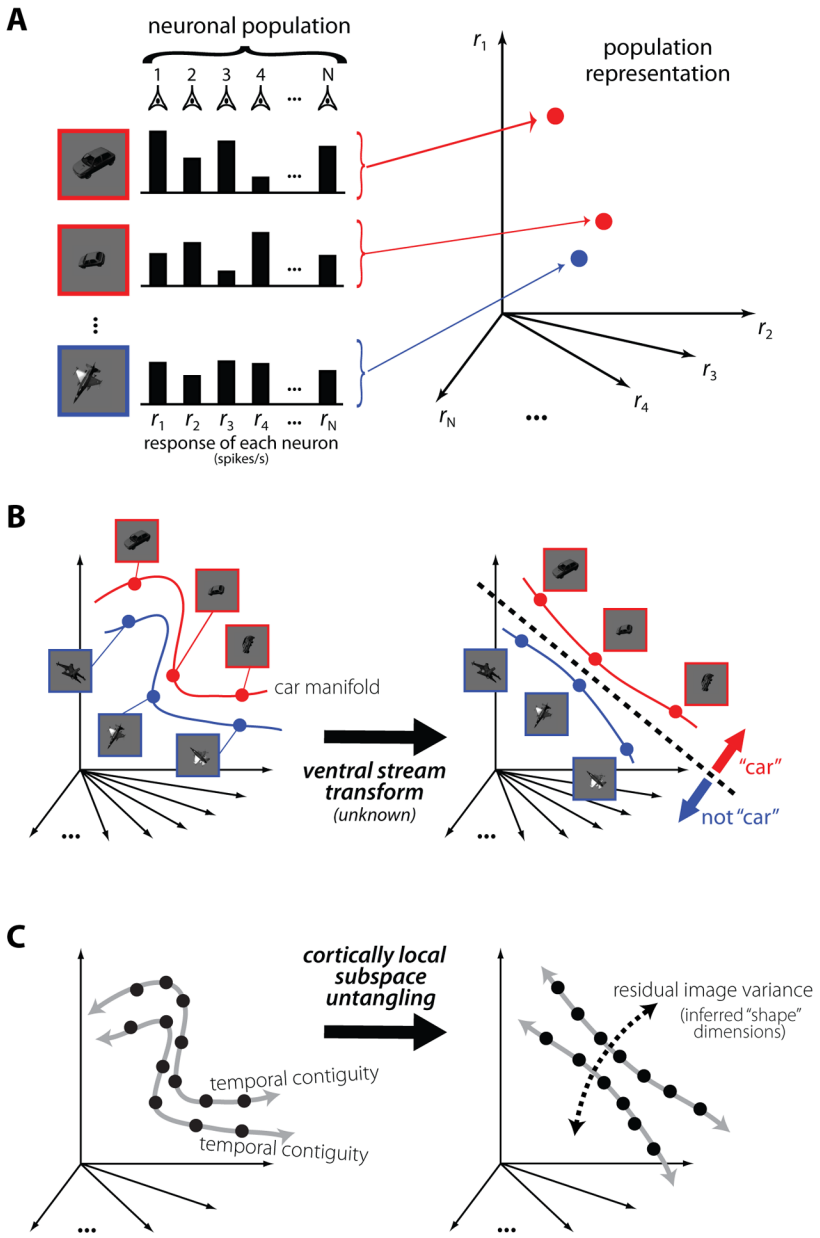

Figure 2. Untangling object representations

(**A**) The response pattern of a population of visual neurons (e.g., retinal ganglion cells) to each image (three images shown) is a point in a very high dimensional space where each axis is the response level of each neuron. (**B**) All possible identity-preserving transformations of an object will form a low-dimensional manifold of points in the population vector space, i.e., a continuous surface (represented here, for simplicity, as a one-dimensional trajectory; see red and blue lines). Neuronal populations in early visual areas (retinal ganglion cells, LGN, V1) contain object identity manifolds that are highly curved and tangled together (see red and blue manifolds in left panel). The solution to the recognition problem is conceptualized as a series of successive re-representations along the ventral stream (black arrow) to a new population representation (IT) that allows easy separation of one namable object's manifold (e.g., a car; see red manifold) from all other object identity manifolds (of which the blue manifold is just one example). Geometrically,

this amounts to remapping the visual images so that the resulting object manifolds can be separated by a simple weighted summation rule (i.e. a hyperplane, see black dashed line; see (DiCarlo and Cox, 2007). (**C**) The vast majority of naturally experienced images are not accompanied with labels (e.g. “car”, “plane”), and are thus shown as black points. However, images arising from the same source (e.g. edge, object) tend to be nearby in time (gray arrows). Recent evidence shows the ventral stream uses that implicit temporal contiguity instruction to build IT neuronal tolerance, and we speculate that this is due to an unsupervised learning strategy termed cortical local subspace untangling (see text). Note that, under this hypothetical strategy, “shape coding” is not the explicit goal -- instead, “shape” information emerges as the residual natural image variation that is not specified by naturally occurring temporal contiguity cues.

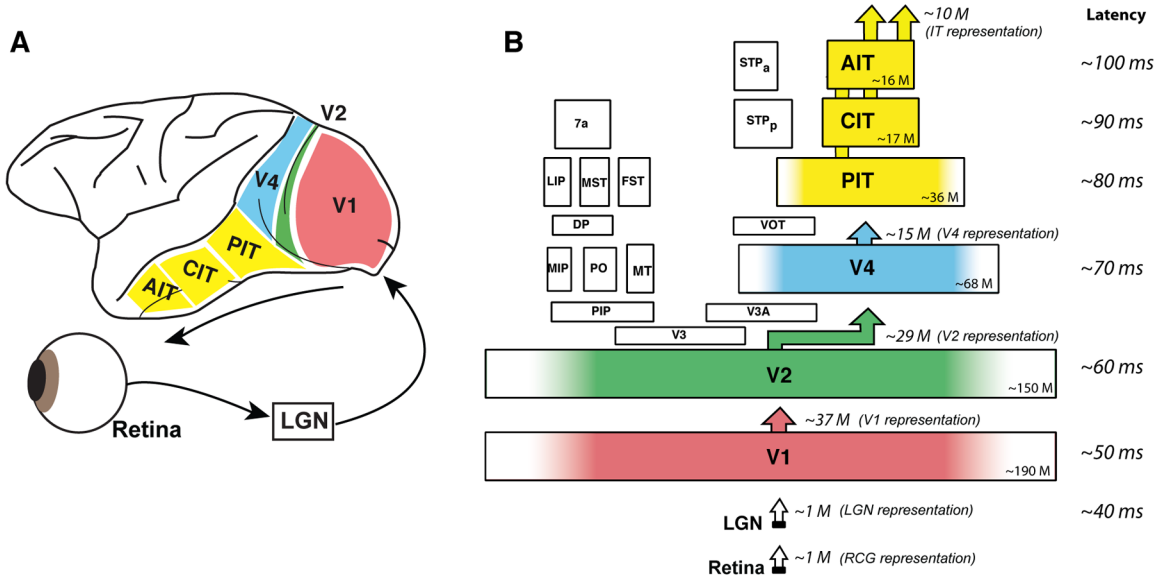

Figure 3. The ventral visual pathway

(**A**) Ventral stream cortical area locations in the macaque monkey brain, and flow of visual information from the retina. (**B**) Each area is plotted so that its size is proportional to its cortical surface area (Felleman and Van Essen, 1991). Approximate total number of neuron (both hemispheres) is shown in the corner of each area (M = million). The approximate dimensionality of each representation (number of projection neurons) is shown above each area, based on neuronal densities (Collins et al., 2010), layer 2/3 neuronal fraction (O'Kusky and Colonnier, 1982), and portion (color) dedicated to processing the central 10 deg of the visual field (Brewer et al., 2002). Approximate median response latency is listed on the right (Nowak and Bullier, 1997; Schmolesky et al., 1998).

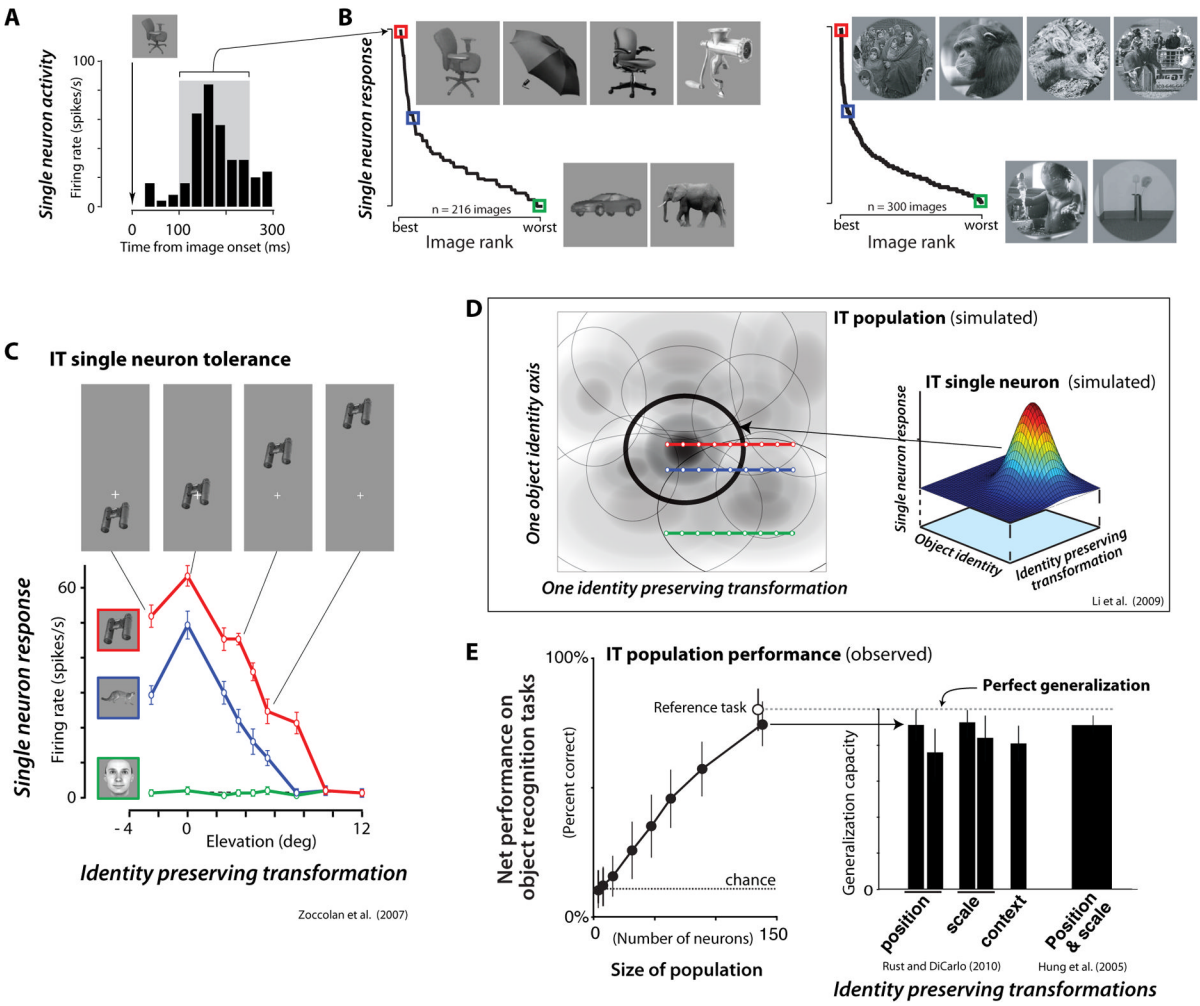

Figure 4. IT single unit properties and their relationship to population performance

(**A**) Post stimulus spike histogram from an example IT neuron to one object image (a chair) that was the most effective among 213 tested object images (Zoccolan et al., 2007). (**B**) Left: The mean responses of the same IT neuron to each of 213 object images (based on spike rate in the gray time window in A). Object images are ranked according to their effectiveness in driving the neuron. As is typical, the neuron responded strongly to ~10% of objects images (four example images of nearly equal effectiveness are shown) and was suppressed below background rate by other objects (two example images shown), with no obvious indication of what critical features triggered or suppressed its firing. Colors indicate highly-effective (red), medium-effective (blue) and poorly-effective (green) images. Right: Data from a second study (new IT neuron) using natural images patches to illustrate the same point (Rust and DiCarlo, unpublished). (**C**) Response profiles from an example IT neuron obtained by varying the position (elevation) of three objects with high (red), medium (blue), and (low) effectiveness. While response magnitude is not preserved, the rank-order object identity preference is maintained along the entire tested range of tested positions. (**D**) To explain data in C, each IT neuron (right panel) is conceptualized as having joint, separable tuning for shape (identity) variables and for identity-preserving variables (e.g. position). If a population of such IT neurons tiles that space of variables (left panel), the resulting population representation conveys untangled object identity manifolds (Fig. 2B, right), while still conveying information about other variables such as position, size, etc. (Li et al., 2009). (**E**)

Direct tests of untangled object identity manifolds consist of using simple decoders (e.g. linear classifiers) to measure the cross-validated population performance on categorization tasks (adapted from (Hung et al., 2005; Rust and DiCarlo, 2010). Performance magnitude approaches ceiling level with only a few hundred neurons (left panel), and the same population decode gives nearly perfect generalization across moderate changes in position (1.5 deg and 3 deg shifts), scale (0.5x/2x and 0.33x/3x), and context (right panel), which is consistent with previous work (Hung et al., 2005); right bar) and with the simulations in (D).

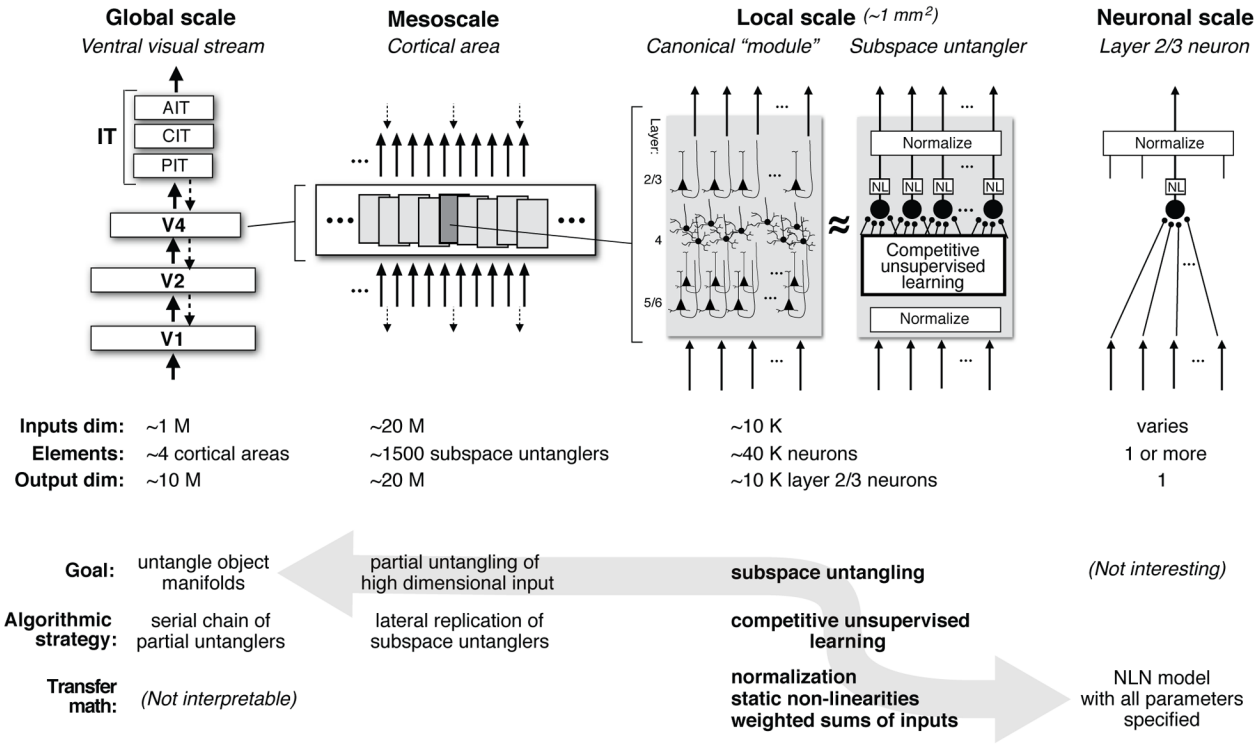

Figure 5. Abstraction layers and their potential links

Here we highlight four potential abstraction layers (organized by anatomical spatial scale), and the approximate number of inputs, outputs, and elemental sub-units at each level of abstraction (M=million, K= thousand). We suggest possible computational goals (what is the "job" of each level of abstraction?), algorithmic strategies (how might it carry out that job?), and transfer function elements (mathematical forms to implement the algorithm). We raise the possibility (gray arrow) that local cortical networks termed "subspace untanglers" are a useful level of abstraction to connect math that captures the transfer functions emulated by cortical circuits (right most panel), to the most elemental type of population transformation needed to build good object representation (see Fig. 2C), and ultimately to full untangling of object identity manifolds (as hypothesized here).

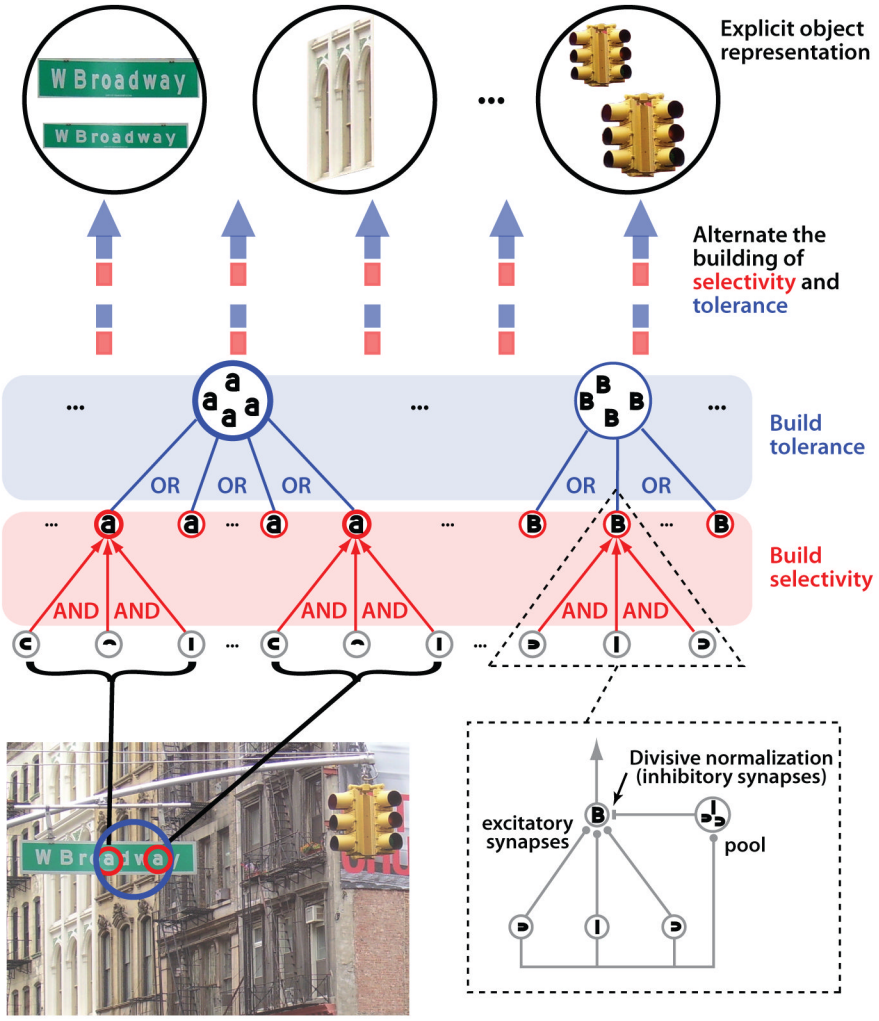

Figure 6. Serial-chain discriminative models of object recognition

A class of biologically-inspired models of object recognition aims to achieve a gradual untangling of object manifolds by stacking layers of neuronal units in a largely feedforward hierarchy. In this example, units in each layer process their inputs using either AND-like (see red units) and OR-like (e.g. "MAX", see blue units) operations, and those operations are applied in parallel in alternating layers. The AND-like operation constructs some tuning for combinations of visual features (e.g. simple cells in V1), and the OR-like operation constructs some tolerance to changes in (e.g.) position and size by pooling over AND-like units with identical feature tuning, but having receptive fields with slightly different retinal locations and sizes. This can produce a gradual increase of the tolerance to variation in object appearance along the hierarchy (e.g. (Fukushima, 1980; Riesenhuber and Poggio, 1999b; Serre et al., 2007a). AND-like operations and OR-like operations can each be formulated (Kouh and Poggio, 2008) as a variant of a standard LN neuronal model with nonlinear gain control mechanisms (e.g. a type of NLN model, see dashed frame).