Reverse Engineering Object Recognition

by

David Daniel Cox

A.B. Biology and Psychology
Harvard University, 2000

Submitted to the Department of Brain and Cognitive Sciences in
partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN NEUROSCIENCE
at the
Massachusetts Institute of Technology

May 2007

© 2007 David Daniel Cox. All rights reserved.

The author hereby grants to MIT permission to
reproduce and to distribute publicly paper and electronic
copies of this thesis document in whole or in part in
an medium now known or hereafter created.

Signature of Author: _____

Department of Brain and Cognitive Sciences
May 17, 2007

Certified by: _____

James J. DiCarlo M.D., Ph.D.
Associate Professor of Neuroscience

Accepted by: _____

Matthew Wilson, Ph.D.
Chair, Graduate Committee
Professor of Neuroscience

Reverse Engineering Object Recognition

by

David Daniel Cox

A.B. Biology and Psychology
Harvard University, 2000

Submitted to the Department of Brain and Cognitive Sciences
on April 24th in partial fulfillment of the requirements
for the degree of Doctor of Philosophy at the
Massachusetts Institute of Technology

Abstract

Any given object in the world can cast an effectively infinite number of different images onto the retina, depending on its position relative to the viewer, the configuration of light sources, and the presence of other objects in the visual field. In spite of this, primates can robustly recognize a multitude of objects in a fraction of a second, with no apparent effort. The computational mechanisms underlying these amazing abilities are poorly understood. This thesis presents a collection of work from human psychophysics, monkey electrophysiology, and computational modelling in an effort to reverse-engineer the key computational components that enable this amazing ability in the primate visual system.

Thesis Superviser: James J. DiCarlo
Title: Assistant Professor of Neuroscience

Acknowledgements

The work presented here would not have been possible without the support and encouragement of many individuals.

It is often said that we stand on the shoulders of those who came before us in science, but the same is also true of our contemporaries and peers. As such, I would like to specially thank my coauthors on the papers that comprise this thesis: Philip Meier and Nadja Oertelt (Chapter 2); Davide Zoccolan (Chapter 3); Nicolas Pinto (Chapter 5), and of course, James DiCarlo (throughout). Thanks as well to the members of the DiCarlo lab, whose discussions and help have been invaluable.

I would also like to thank the members of my thesis committee, Pawan Sinha, Nancy Kanwisher, and John Maunsell, for their helpful discussions and guidance. I would also be remiss if I did not thank my advisor, Jim DiCarlo, whose mentorship and support made this work possible.

Finally, I'd like to thank my family for their unwavering support: my parents, Jerry and Elaine, my brothers Michael and Brian, and especially my wife, Jarasa.

Reverse Engineering Object Recognition
Table of Contents

Preface

We recognize visual objects with such ease that it is easy to overlook what an impressive computational feat this represents. Any given object in the world can cast an effectively infinite number of different images onto the retina, depending on its position relative to the viewer, the configuration of light sources, and the presence of other objects in the visual field. Further compounding the problem, there are at least hundreds of thousands of distinct object classes that we must recognize, and we must also be able to deal with fundamentally novel objects that we have never seen before. In spite of these issues, the visual system is able to robustly identify objects, all in a fraction of a second.

Neurons in the inferotemporal cortex (IT) of primates – the finally exclusively visual area in the ventral temporal visual pathway – respond selectively to complex visual objects while at the same time tolerating a range of variation in the objects' retinal image. By the time visual information passes through IT, many of the "hard" problems in object recognition have somehow been solved. IT thus represents an attractive target for reverse engineering object recognition

However, our understanding of how IT works is still in its infancy. From a computational perspective, we lack models that can produce realistic or useful recognition behavior, except in the context of "toy" problems. At the same time, from a neurophysiology perspective, there exists no coherent plausible description of what visual features IT neurons are tuned for, nor can we generate models that will predict the responses of IT. This is not to say that the problem is insoluble; rather it is to say that much work remains to be done.

This thesis seeks to clearly identity some of the fundamental "core" problems of object recognition and to provide some basic empirical footholds in tackling these problems, using techniques spanning from human psychophysics to primate neurophysiology to computational modeling.

Chapter 1 (Untangling Object Recognition) presents a perspective on the true core challenges of object recognition and serves as an introduction to the rest of the thesis. In particular, we introduce the notion of manifold "tangling" as a way of conceptualizing why the problem is hard and suggesting what kinds of computations might be useful in solving the problem. We also lay out a potential path forward, and relate this path to the remainder of this thesis.

Chapter 2 ("Breaking" Position-Invariant Object Recognition) describes one effort using human psychophysics to test the hypothesis that time may play in learning visual invariance.

Chapters 3 & 4 (Multiple Object Response Normalization in Monkey Inferotemporal Cortex and Can Inferotemporal Cortex Simultaneously Represent Multiple Objects?) describe efforts to understand how single unit and population responses in inferotemporal cortex behave when multiple objects are present. In particular, we use approaches outlined in Chapter 1 to ask what kinds of information can be extracted from an IT representation.

Finally, **Chapter 5** ("Why is Natural Vision Hard?") presents an early attempt to understand what are the primary challenges in the construction of artificial visual object recognition systems. In particular, we show that currently available "gold standard" object recognition test sets do not properly exercise those aspects of the problem that are truly difficult.

Chapter 1:
Untangling Object Recognition

The chapter was submitted to Trends in Cognitive Sciences in February 2007

Any given object can cast an infinite number of different images onto the retina, depending on its position and pose relative to the viewer, the configuration of light sources, and the presence of other objects in the visual field. In spite of this, primates can recognize a multitude of objects, each in a fraction of a second, with no apparent effort. The computational mechanisms embedded in the brain that enable this invariant object recognition ability are poorly understood, and their elucidation would have broad implications for perceptual and cognitive science. Here, we present a graphical perspective on invariant object recognition, drawing on key ideas from the neurophysiology and computational literature. The perspective is intended to foster insight into the computational crux of the problem – what we term object "tangling"— and to illustrate what solutions might look like. We then set neuronal data from the primate visual system in that perspective and argue that it has achieved a potentially optimal solution where strict invariance is not necessarily the goal. Finally, we speculate on the step-wise operations the visual system may use to achieve this solution, and outline steps towards understanding these mechanisms.

Introduction

Humans are highly dependent on visual object recognition: our daily activities rely on accurate and rapid identifications of objects in our visual environment. The apparent ease of object recognition belies the magnitude of this feat – we effortlessly recognize objects from among tens of thousands of possibilities, and we do this within a fraction of a second, in spite of tremendous variation in the appearance of each one. Understanding the brain representations and mechanisms that underlie this ability would be a landmark achievement in neuroscience.

Object recognition is computationally difficult for a number of reasons, but the most fundamental reason is that any individual object can produce an infinite set of different images on the retina, due to variation in (e.g.) object position, scale, pose, illumination, and the presence of visual clutter. Indeed, although we typically see each object many times, virtually every image on our retina is different from all previous images. Thus, the key computational challenge of object recognition is extracting object identity, in spite of large amounts of image variation (e.g. Ullman,

1996; Ashbridge and Perrett, 1998; Edelman, 1999; Riesenhuber and Poggio, 1999; Rolls, 2000). Although a large number of computational efforts have attacked this so-called "invariance problem" (e.g. Biederman, 1987; Olshausen et al., 1993; Bengio et al., 1995; Wallis and Rolls, 1997; Edelman, 1999; Riesenhuber and Poggio, 1999; Ullman and Soloviev, 1999; Arathorn, 2002; Yuille and Kersten, 2006; Serre et al., 2007), a robust, real-world machine solution still evades us, and we are far from a satisfying understanding of how the problem is solved in the brain. We believe that these two achievements will be accomplished nearly simultaneously by an approach that is savvy to both the computational issues and the biological clues and constraints.

In this opinion piece, we use a graphical perspective to provide intuition about the invariance problem, show that the primate ventral visual processing stream produces a particularly effective solution in inferotemporal cortex (IT), give our opinion on how the ventral stream approaches the problem, and propose directions for future research. To do this, we bring together some of the most important ideas from computation and neurophysiology. Individually, several of these ideas have been raised previously by others in related contexts. However, because it is easy for one to get lost in the very large sea of previous studies and ideas, the central contribution of this manuscript is to clear the table, to bring forth what we believe to be the key ideas in the context of what is known about the primate brain, and to pull those threads together into a coherent framework. Along the way, we show that some ideas and approaches that may appear important, are only tangential to, or even distract from, the goal of understanding invariant object recognition.

What is object recognition?

We define invariant object recognition (from here on, simply "object recognition") as the ability to: 1) accurately discriminate each named object ("identification") or set of objects ("categorization") from all other possible objects, materials, textures, etc. found in a visual world, and 2) do this over a range of transformations of the retinal image of that object ("identity preserving transformations").

Of course, vision encompasses many disparate challenges besides object recognition, such as materials and texture recognition, object similarity estimation, object segmentation, object tracking and trajectory prediction, etc. Although some of these challenges may interact with object recognition either as underlying requirements for object recognition or as natural extensions of object recognition, exploring those possible relationships is not our goal. Instead, we aim to see how far a clear focus on the problem of object recognition will take us. For even more clarity, we concentrate on what we believe to be the core of the brain's recognition system -– the ability to rapidly report object identity or category after just a single brief glimpse of visual input (<300 ms; Potter, 1976; Thorpe et al., 1996).

What computational processes must underlie object recognition?

The problem of object recognition is fundamentally a problem of data representation and re-representation ((also see Marr, 1982; Johnson et al., 1995)). The visual system takes incoming information from the retina and transforms ("re-represents") this information into a form that

can be easily used for a variety of tasks (object identification, object categorization, etc.). By "representation" we simply mean the visual information in the activity of a population of neurons (though others have used this word to mean more Edelman, 1999). Below, we will see that the activity of the population of neurons in IT cortex is a particularly good representation of object identity. However, first we must motivate our focus on representation.

When a subject correctly solves a perceptual task, such as recognition, the subject must be using some internal neuronal representation of part or all of the visual scene to make a decision (e.g. Johnson, 1980; Ashby and Gott, 1988): "Is object A present or not?" Computationally, the brain must apply a decision function (Johnson, 1980) to divide an underlying neuronal representational space into regions where object A is present and regions where it is not (Fig. 1b; one function for each object to be potentially reported). Given that brains compute with neurons, the subject must have neurons somewhere in its nervous system -- "read-out" neurons -- that can successfully report if object A was present (Barlow, 1995). Of course, there are many relevant mechanistic details such as: how many such neurons are involved in computing the decision, where are they in the brain, is their operation fixed or dynamically created with the task at hand, and how do they code choices in their spiking output? But these are not the central computational issues of object recognition. The central computational issues are: 1) what is the format of the representation used to support the decision (the substrate on which the decision functions directly operate), and 2) what kinds of decision functions (i.e. "read-out" tools) are applied to that representation?

These two central computational issues are two sides of the same coin. For example, one can cast object recognition to be the problem of finding very complex decision functions (highly-nonlinear) that operate on the retinal image representation. This is like trying to swallow the recognition problem whole. Alternatively, one can cast the recognition problem as one of finding operations that gradually transform that same retinal representation into a new form of representation, followed by the application of relatively simple decision functions (e.g. linear classifiers Duda et al., 2001) to that new representation. From a computational perspective, the difference is largely terminology, but we and others (e.g. Johnson, 1980; Hung et al., 2005) argue that the latter viewpoint is more productive because it starts to take the problem apart in a way that is consistent with what we know about the architecture and response properties of the ventral visual stream, and because such simple decision functions can be easily implemented in a single, biologically-plausible neuronal processing step (a thresholded sum over weighted synapses). This point of view also meshes well with the conventional wisdom in the field of pattern recognition -- choice of representation is often more important than the "strength" of the classifier used. As show below, a variety of recognition tasks can be solved in IT population responses using simple, linear classifiers (Hung et al., 2005), suggesting that our decision to focus on such operations is not unreasonable. That is, even if the brain does have access to substantially more complex decision rules, many real problems in recognition can be solved without invoking greater complexity (in addition, more complex decision functions would generally also benefit from better linear separability). Finally, even from this viewpoint, one is still completely free to consider the possibility that the computations to implement "representation" are not substantially different from those applied during "classification" using that representation. Thus, with little loss of generality and only minimal assumptions of underlying neuronal mechanisms, below we use simple (linear) decision functions to examine the usefulness of representations that might underlie object

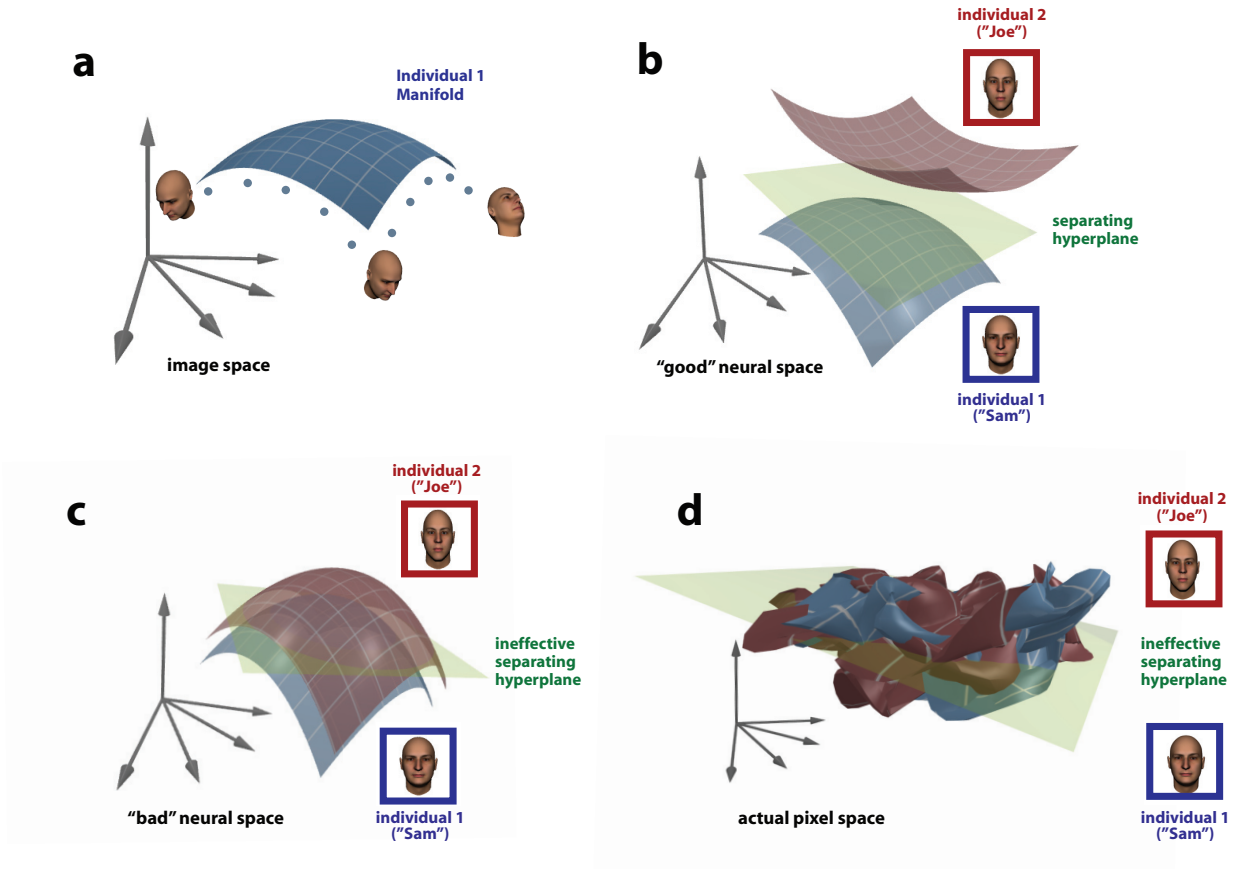


Figure 1. Conceptual illustration of object tangling.

In a neuronal population space, each cardinal axis is the activity of a single neuron (e.g. a retinal ganglion cell). Although such high-dimensional spaces cannot be visualized, the three-dimensional version portrayed here provides the fundamental intuitions. a) A given image of a single object (in this case, a particular face) is represented as a single point in retinal image space. We conceptually illustrate what happens as the face is gradually transformed in pose in the external world (relative to the eye of the viewer) and projected onto the viewer's retina. Because only two dimensions of pose are varied here, the point representing the object travels through a two-dimensional space (the blue surface). In retinal image space, such identity-preserving transformations cause the point not to move in a straight line, but along a curved path. Thus, the two-dimensional space is not a flat plane, but a curved surface called a manifold. A key point is that all possible images of this particular object are contained within its manifold. b) The manifolds of two objects (two faces, red and blue) are shown in a common neuronal population space. In this case, a flat decision plane (a biologically-plausible decision rule) can be drawn cleanly between them. If the world consisted only of these two objects over this amount of variation, this neuronal representation is "good" for supporting visual recognition. c) In this case, the two object manifolds are intertwined, or tangled. The decision plane can no longer separates the two manifolds no matter how the plane is tipped or translated. d) The pixel (retinal-like) manifolds generated from actual models of faces, undergoing two types of pose changes. The 3D display axes were chosen by finding the best projections that separate identity, pose azimuth and pose elevation as the two manifolds as the two faces were varied over pose, position, size and illumination. Clearly, even in this simple two-object world, the object manifolds are hopelessly tangled in the retinal image, and this is fundamentally due to natural variation, not ambiguity in the image.

recognition.

Why is object recognition hard? Object manifold tangling

Object recognition is hard because useful forms of visual representation are not easy to build. In part, this is because we do not have good intuition about visual representations because vision operates in very high dimensional spaces. Our eyes fixate the world in ~300 ms intervals before moving on to a new location. During this brief glimpse, a visual image is projected into the eye, transduced by ~100 million photoreceptors arrayed along the retina, and conveyed to the brain in the spiking activity pattern of ~1 million retinal ganglion cells (Wandell, 1995). Such a representation can be conceptualized as a high-dimensional extension of a simple three-dimensional Cartesian space in which each axis of the space is the response of one retinal ganglion cell (e.g. Roweis and Saul, 2000; Tenenbaum et al., 2000; see Fig. 1). Ignoring temporal information and measuring each neuron's response to each glimpse as its mean spiking rate, each possible image projected into the eye is one point in a ~1 million dimensional retinal ganglion cell representation.

To gain intuition about high-dimensional visual representations, note that, within this immense retinal representation space, different encounters with the same physical object lie in contiguous regions. For example, consider just one glimpse of a particular face. That single glimpse of that face, in exactly that position, scale, pose, lighting, and background produces just one pattern of activity on your retina – it is just one point in the retinal image space (note that we ignore internal "noise," which would blur the point around a true mean, but this is not fundamental to our arguments). Now imagine all the possible retinal images that that particular face could ever possibly produce (e.g. due to changes in its pose, position, size, etc.), and the corresponding set of points in the retinal image space. Together, that set of potential data points arises from a continuous, low-dimensional, curved surface inside the retinal image space called an object "manifold" (e.g. see Edelman, 1999; Roweis and Saul, 2000; Tenenbaum et al., 2000). Different objects have different manifolds (see Fig. 1b-d).

Given this framework, we start with a simple world of just two possible objects (Joe and Sam, see Figure 1), to graphically show the difference between a "good" and "bad" representation for directly supporting object recognition. The representation in Figure 1b is good: it is easy to determine if Joe is present, in spite of pose variation, by simply placing the linear decision function (i.e. a plane) between Joe's manifold and the rest of the other potential images in the visual world (just images of Sam in this case). In contrast, the representation in Figure 1c is not well suited to recognition, because, in this representational space, the object manifolds are "tangled," such that it is impossible to reliably separate Joe from the rest of the visual world with a linear decision function.

Are object manifolds tangled this way in real life? Figure 1d shows actual 14,400-dimensional pixel data (120x120 images) for the two face objects in the presence of mild variation in their pose, position, scale, and lighting, projected into a three dimensional space with linear axes chosen to maximally separate the two objects. Even in this simple example that only exercises a fraction of typical real-world variation, the manifolds are hopelessly tangled. This demonstration

graphically reveals why the retinal representation cannot directly support object recognition -- because each object's manifold is hopelessly tangled together with other object manifolds.

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

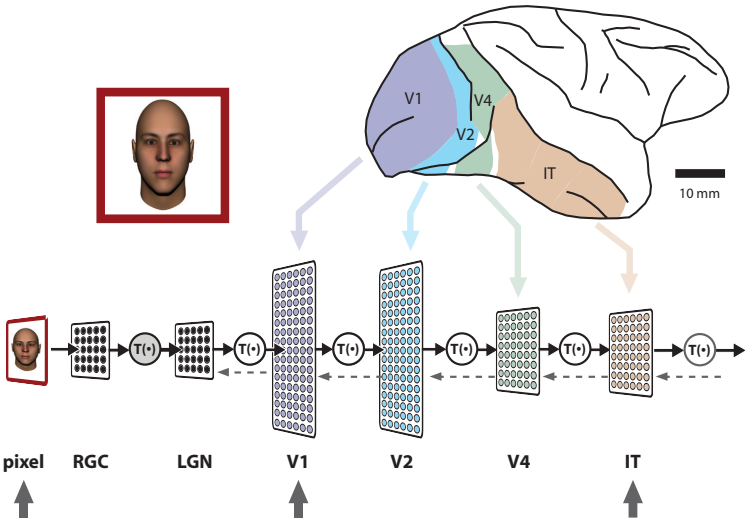


Figure 2. Neuronal populations along the ventral visual processing stream

Although we seek to ultimately understand how object recognition is accomplished by the human brain, the rhesus monkey is our current best model system. Like humans, this species has high visual acuity, relies heavily on vision (~50% of macaque neocortex is devoted to vision), and easily performs visual recognition tasks. Moreover, many visual areas of the rhesus monkey have been well mapped and are hierarchically organized 71. A battery of previous work tells us that the ventral visual stream is important for complex object discrimination (Tanaka 1996, Logothetis & Sheinberg 1996, Miyashita 1993, Ungerleider & Mishkin 1982, Afraz et al. 2006). We show a lateral schematic of a rhesus monkey brain, with the areas of the ventral visual stream colored (adapted from Felleman & Van Essen 1991). Lower panels illustrate neuronal populations in early visual areas and at successively higher stages across the ventral visual stream. A given pattern of photons in the world (here a face) is transduced into neuronal activity at the retina. Here we conceptualize each processing stage of the ventral stream as a new population representation (each population is known to span ~10 deg of central vision; for simplicity they are each shown as equal in size). A single image can be considered as a point in each ventral stream population representation (red dot, see Fig. 1). Arrows indicate the direction of visual information flow based on neuronal latency, but this does not preclude fast feedback both within and between areas. The population representations for the retina, V1, and AIT are considered in Figure 1e and 3a,b.

One way of viewing the overarching goal of the brain's object recognition machinery, then, is as a transformation from visual representations that are easy to build (e.g. center-surround filters in the retina), but are not easily decoded (as in Fig. 1c,d), into representations that we do not yet know how to build, but are easily decoded (e.g. IT; Fig. 1b, see below). Although the idea of representational transformation has been stated under a number of guises (2 1/2D sketch, feature selection, etc. Marr, 1982; Johnson et al., 1995; Duda et al., 2001), we argue below that the untangling perspective goes further, by suggesting the kinds of transformations the ventral visual system should perform if its goal is to accomplish good representation for object recognition. But first we look at the primate ventral visual stream from this untangling perspective.

The ventral visual stream transformation untangles object manifolds

In humans and other primates, the key information processing to support visual recognition likely takes place along the ventral visual stream (for review see Logothetis and Sheinberg, 1996; Tanaka, 1996). We, like some others (e.g. Felleman and Van Essen, 1991; Riesenhuber and Poggio, 1999), consider this stream to be a progressive series of visual re-representation, from V1 to V2 to V4 to IT (see Fig. 2). Since the pioneering studies of Gross and colleagues (Gross et al., 1972; Desimone et al., 1984), a wealth of work has shown that single neurons at the highest level of the monkey ventral visual stream – the inferotemporal cortex (IT) – display spiking responses that are likely useful for object recognition (Logothetis et al., 1995; Tanaka, 1996; Rolls, 2000). Specifically, many individual IT neurons respond selectively to particular classes of objects, such as faces (Perrett et al., 1982; Desimone et al., 1984; Tsao et al., 2006) or other complex shapes (Ungerleider and Mishkin, 1982; Desimone et al., 1984; Miyashita, 1993; Logothetis et al., 1995; Logothetis and Pauls, 1995; Logothetis and Sheinberg, 1996; Tanaka, 1996; Sheinberg and Logothetis, 2001), yet show tolerance to limited changes in object position and size (Schwartz et al., 1983; Sary et al., 1993; Tovée et al., 1994; Ito et al., 1995; Logothetis et al., 1995; Op de Beeck and Vogels, 2000; DiCarlo and Maunsell, 2003), pose (Logothetis et al., 1995; Booth and Rolls, 1998), illumination (Vogels and Biederman, 2002) and low level shape cues (Sary et al., 1993).

How does one use the response of these individual ventral stream neurons (e.g. IT, above) to gain insight into object manifold untangling in the brain? To do this, our group has focused on characterizing the initial wave of neuronal population "images" that are successively produced along the ventral visual stream as the retinal image is transformed and re-represented on its way to IT (see Fig. 2). For example, looking at the end of the stream, we and our collaborators found that simple linear classifiers can rapidly (within <300 ms from image onset) and accurately decide an object's category from an IT population of ~200 neurons, despite object position and size changes (Hung et al., 2005). Because the type of classifier did not much matter, and the same classifiers fail at the same task when applied to a simulated V1 population of equal size (Hung et al., 2005), this performance is not due to the classifiers themselves, but to the powerful form of visual representation conveyed by IT. This shows that, compared with early visual representations, object manifolds are less tangled in the IT population representation.

To show that this untangling happens as suggested in Figure 1b, Figure 3 illustrates object manifolds in primary visual cortex (V1) and IT. These are the manifolds of the faces of Sam and Joe from Figure 1d (retina-like representation), but now shown re-represented in cortical population

space. To generate these, we took populations of simulated response functions from previous single unit work in V1 (e.g. Hubel and Wiesel, 1977; Ringach, 2002) and IT (e.g Ito et al., 1995; Logothetis et al., 1995; Op de Beeck and Vogels, 2000), and applied them to all the images of Joe and Sam. This reveals that the V1 representation, like the retinal representation, still contains highly curved, tangled object manifolds (Fig. 3a), while the same object manifolds are flattened and untangled in the IT representation (Fig. 3b). It is easy to see that, from the point of view of (potentially simple) downstream decision neurons, the retina and V1 representations are not in a good format to separate Joe from the rest of the world, while the IT representation is. In sum, the experimental evidence suggests that the ventral stream transformation (culminating in IT) solves object recognition by untangling object manifolds. For each visual image striking the eye, this total transformation happens gradually (i.e. step-wise transformations along the cortical stages), but rapidly (i.e. <100 ms from V1 to IT, ~20 ms per cortical stage). But what is this transformation? That is, how does the ventral stream do this?

How does the ventral visual stream untangle object manifolds?

We do not yet know the answer to this question. As a start, Hubel and Weisel's (Hubel and Wiesel, 1977) observation that V1 simple cells are shape selective (orientation) and V1 complex cells create some tolerance to identity-preserving transformations (esp. position) has been computationally implemented and extended to higher cortical levels including "IT" (Fukushima, 1980; Riesenhuber and Poggio, 1999; Serre et al., 2007). But, beyond this early insight, systems neuroscience has not provided a breakthrough as to how the ventral visual stream constructs key aspects of high-level population responses (e.g. V4 or IT).

Some important neurophysiological effort has focused on characterizing the tolerance of individual IT neurons to variation in each object's image (e.g. Tovée et al., 1994; Ito et al., 1995; Logothetis et al., 1995; Op de Beeck and Vogels, 2000; Vogels and Biederman, 2002; DiCarlo and Maunsell, 2003; Zoccolan et al., 2005), which is central to object tangling. However, much more effort has been aimed at understanding effects of behavioral states (e.g. task, attention Moran and Desimone, 1985; Sato, 1988; Chelazzi et al., 1993; Motter, 1994; Maunsell, 1995; Vogels et al., 1995; McAdams and Maunsell, 1999; DiCarlo and Maunsell, 2000; Naya et al., 2003; Reynolds and Desimone, 2003; Suzuki et al., 2006). While such studies have made great progress in showing the ways in which neuronal responses are modulated by behavioral state, they side-step the key problem of untangling, because these effects can be measured without a deep understanding of the format of visual representation in the brain area examined.

Substantial effort has also been aimed at understanding the features or shape dimensions in visual images to which V4 and IT neurons are most sensitive (e.g. Gallant et al., 1993; Tanaka, 1996; Pasupathy and Connor, 2001; Tsunoda, 2001; Pollen et al., 2002; Kayaert et al., 2003; Brincat and Connor, 2004; Yamane et al., 2006). Such studies are important in that they help define the feature complexity of neuronal tuning at each level of the ventral stream, which is indirectly related to object tangling (because object tangling implicitly assumes a representation that is untangled with respect to "objects" or, at least, conjunctions of features). Following on this line, current, ambitious approaches to fully understand the response functions of individual neurons (i.e. the non-linear operators on the visual image) would, if successful, lead to a full understanding of

visual representation and thus bring an implicit understanding of object recognition. However, given the enormity of this task, it is not surprising that progress has been slow.

In contrast, the object untangling perspective presented here leads to a complementary, but qualitatively different approach. First, it shifts one from thinking about ideal single unit response properties in IT (Barlow, 1995; Gross, 2002) – which is akin to studying individual feathers to understand flight 17– to thinking about ideal formats of population representation with the computational goals of the behavioral task clearly considered (see Fig. 3b vs. 3c) (Salinas, 2006). Second, it suggests the immediate goal of determining how well neuronal representations along

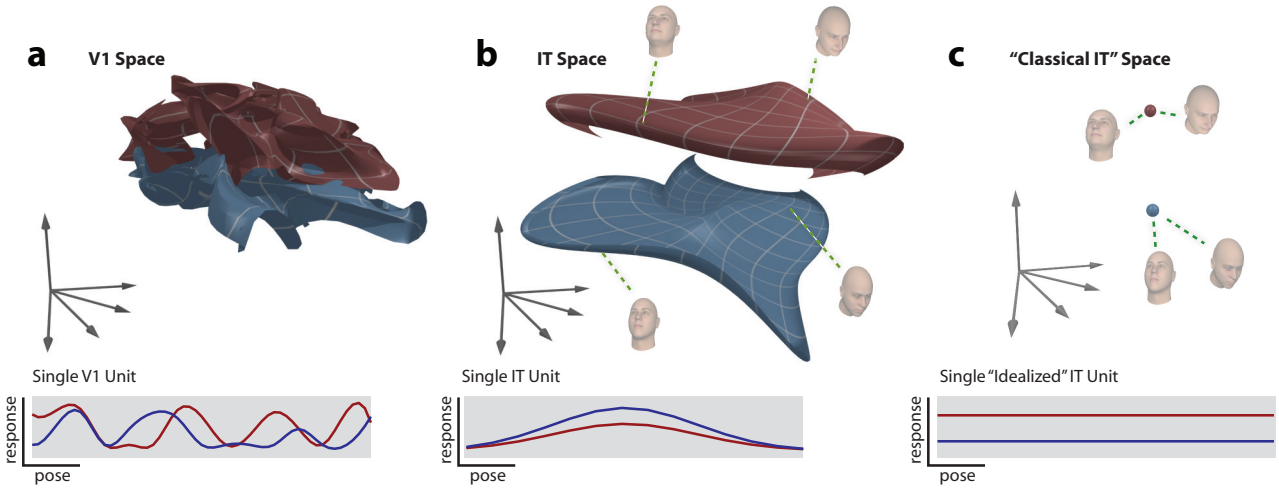


Figure 3. Untangling of object manifolds along the ventral visual stream

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

the ventral stream have untangled object manifolds, and it shows how to quantitatively measure untangling (see linear classifiers above, Fig. 1). Third, this perspective points to new ways to compare computational models to neuronal data. The evaluation of computational models at the single unit level is problematic because such comparisons are typically under-constrained – that is, to make a comparison, we must typically fit a large number of model parameters to a relatively small amount of neuronal data, and the flexibility of the model often overwhelms the constraints placed on it by the data. The untangling perspective suggests that such comparisons might be more meaningful at the population level (e.g. one could make predictions of degree of untangling at different levels of the ventral stream). Fourth, it suggests a clear focus on the causes of tangling – identity-preserving transformations – rather than the traditional focus on 'shape' or 'features'. Indeed, because we fundamentally lack an understanding of the dimensionality of 'shape' or how to manipulate those dimensions experimentally, we speculate that computational/experimental approaches that focus on tolerance across identity-preserving transformations while simply preserving/measuring sensitivity to other real-world image variation (e.g. learned to flatten object manifolds, see below) will be vastly more tractable. Finally, this perspective steers experimental effort toward testing hypothetical mechanisms that might underlie untangling (e.g. Wallis and Bulthoff, 2001; Cox et al., 2005), and it steers complimentary computational effort toward finding new, biologically-plausible algorithms that might gradually untangle object manifolds (e.g. Wallis and Rolls, 1997; Riesenhuber and Poggio, 1999). We close by discussing our views on this point.

Flattened object manifolds are a good solution

The illustrations in Figure 3 suggest a strategy for building good object representations: if the goal is to untangle manifolds corresponding to different objects, then we seek transforms that "flatten" these manifolds, while preserving selectivity across object identity axes (e.g. "shape" axes). This perspective is partly a restatement of the problem of invariant object recognition, but not an entirely obvious one. For example, the textbook conception of IT suggests a different set of goals for each IT neuron: very high shape selectivity and "invariance" to identity-preserving image transformations. To illustrate how object manifold untangling gives fresh perspective, Figure 3b and c show just two simulated IT populations which have both successfully untangled object identity, but which have very different single unit response properties. In Figure 3c, each single unit has somehow met the textbook ideal of being selective for object identity, yet invariant to identity-preserving transformations. At the IT population level, this results in untangling object manifolds by "collapsing" each manifold to a single point. By comparison, in Figure 3b, every single IT unit has good sensitivity to object identity, but only limited tolerance to object transformation (e.g. position, scale, view) and, by textbook standards, seems less than ideal. However, at the population level, this also results in untangled object manifolds, but in a way that has coordinated (i.e. "flattened"), rather than discarded, information about the transformation variables (pose, position, scale, etc.). This suggests that the IT representation should not only be able to directly support object recognition, it should also directly support tasks such as pose, position, and size estimation, as previously suggested by theorists ((e.g. Edelman, 1999)). Indeed, real IT neurons are not position and size invariant in that they have limited spatial receptive fields (Op de Beeck and Vogels, 2000; DiCarlo and Maunsell, 2003). It is now easy to see that this "limitation" is an advantage, as long as the object manifolds are nearly flat and coordinated

(i.e. if the individual IT neurons have response properties like those in Figure 3b).

Ways the brain might learn to flatten object manifolds

Although the flattening of object manifolds might be partly accomplished by hard-wired transforms (e.g. Fukushima, 1980; Riesenhuber and Poggio, 1999), it has been noted that one could also learn the structure of the manifold from the statistics of natural images (e.g. Roweis and Saul, 2000; Tenenbaum et al., 2000), which would potentially allow a flattened re-representation of the manifold. However, while most previous "manifold learning" efforts have emphasized learning structure in the ambient pixel/retina space in one step, we impose no such requirement. In particular, the transforms need only flatten object manifolds little by little in many successive steps ((this is consistent with physiological data which show that response properties progress gradually along the ventral stream Kobatake and Tanaka, 1994)). The notion of progressive flattening is a matter of both emphasis and substance: there is no need to swallow the entire problem whole – representations can be flattened locally at small scales (both in term of visual space and input dimensionality of the neurons working at the next level) which can ultimately produce flattening at a much larger scale. Indeed, one advantage of the manifold tangling perspective is that it still makes sense at a variety of scales – V1 neurons in a local neighborhood only "see" the world through a small aperture (and thus cannot see whole objects), but they can perform flattening operations with respect to their (relatively restricted) inputs; V2 can do the same on its V1 inputs, and so on (see the discussion of time, below, for thoughts on how neurons can have access to information about manifold degrees of freedom). Thus, we believe that the most fruitful computational algorithms will be those that a visual system (natural or artificial) could apply locally and iteratively at each cortical processing stage (e.g. Heeger et al., 1996) in a largely unsupervised manner (e.g. Einhauser et al., 2005), and that achieve some local object manifold flattening. Even though no single cortical stage or local ensemble within a stage would "understand" its role in this process, we imagine the end result to be globally flattened, coordinated object manifolds with preserved shape selectivity.

In our view, their are three important computational ideas that are consistent with physiology and that, together, may allow flattening to happen. First, the visual system projects incoming information into a higher-dimensional, overcomplete space (e.g. there are ~100 times more V1 neurons than retinal ganglion cells, see Fig. 2). This dimensionality explosion can "spread out" the data into this much larger space. The additional constraint of sparseness can reduce the size of the effective subspace that any given incoming visual image "lives" in and thus make it easier to find projections where object manifolds are flat and separable (see Olshausen and Field, 2004). A second, related idea, is that, at each processing stage, neuronal resources (i.e. neuronal tuning functions on the previous stage) are allocated in a way that matches the distribution of visual information encountered in the real world (e.g. Ullman et al., 2002; Simoncelli, 2003). This would increase the effective over-completeness of visual representations of real-world objects (and thus help flatten object manifolds). Indeed, a variety of biologically plausible algorithms developed in other contexts (e.g. Schwartz and Simoncelli, 2001 , Heeger et al., 1996), may have a yet-to-be-discovered roles in achieving coordinating flattening within local neuronal populations. For example, divisive normalization is a powerful nonlinearity that can literally "bend" representational spaces.

A third, potentially key idea is that time can implicitly supervise manifold flattening. A number of theorists have noticed that the evolution of a retinal image over time provides clues for learning which image changes are identity-preserving transformations and which are not (Foldiak, 1991; Wallis and Rolls, 1997; Ullman and Soloviev, 1999; Edelman and Intrator, 2002; Wiskott and Sejnowski, 2002). In the language of object tangling, this is equivalent to saying that the evolution across time spells out the degrees of freedom of the object manifold. We hypothesize that the ventral stream may use this temporal evolution to achieve progressive "flattening" of object manifolds across each neuronal processing stage. Recent studies in our lab (Cox et al., 2005) and others (Wallis and Bulthoff, 2001) have begun to connect this computational idea with the biological vision, showing that invariant object recognition can be predictably manipulated by the temporal statistics of the environment. We are actively pursing this exciting new direction in the neurophysiology of the ventral stream.

A Path Forward

Although we still have a long way to go to achieve a deep understanding of how the brain accomplishes object recognition, it is a very exciting time to be working on the problem. There is a rapid blurring of lines between traditionally separate fields that each have an interest and each have something unique to bring to the table. We hope that the untangling perspective presented here will facilitate this progress. As we move forward, we propose these overarching guidelines:

• Neuroscience and psychophysical efforts should be aimed at conducting targeted experiments to distinguish among "real" computational models (hypotheses), and developing new methods to obtain such data. **Chapters 2-4** of this thesis are aimed at dissecting the computational underpinnings of invariant object recognition, using a variety of methods. **Chapter 2** explores the role of time as a possible component in invariance learning, using human psychophysics. In the context of manifold tangling, such temporal learning could represent a powerful flattener / untangler. **Chapter 3** explores potential normalization mechanisms in monkey inferotemporal cortex using multiple simultaneous objects. **Chapter 4** examines how inferotemporal cortical population responses might support representation of multiple simultaneously present objects.

• Another important outstanding task is the establishment and construction of concrete specifications and benchmark tests of what problems visual recognition is expected to solve. These tests must directly engage the key challenges of real-world recognition (e.g. outlined here) and thus avoid limited domain heuristics. Practically, these might consist of input / output parings of labeled data (image and video databases) that are freely available to benchmark performance from any field (psychophysics, computation, neurophysiology). Given that we will not have the foresight to develop a complete set of benchmarks, they should naturally grow in difficulty as progress is made. **Chapter 5** of this thesis presents an early step in this direction. In particular, we show some of the pitfalls inherit in choosing a test set, and we present a possible path to avoid these pitfalls.

Chapter 2:
"Breaking" Position-Invariant Object Recognition

The following chapter originally appeared as:

Cox DD, Meier P, Oertelt N, DiCarlo JJ (2005)."Breaking" Position Invariant Object Recognition. Nature Neuroscience 8(9): 1145-1147.

It is often assumed that objects can be recognized irrespective of where they fall on the retina, yet little is known about the mechanisms underlying this ability. By exposing human subjects to an altered world where some objects systematically changed identity during the transient blindness that accompanies eye movements, we induced predictable object confusions across retinal positions, effectively "breaking" position invariance. Thus, position invariance is not a rigid property of vision but is constantly adapting to the statistics of the environment.

Any given object can cast an essentially infinite number of different images on the retina, due to variations in position, scale, view, lighting, and a host of other factors. Nonetheless, humans effortlessly recognize familiar objects in a manner that is largely invariant to these transformations. The ability to identify objects in spite of these transforms is central to human visual object recognition, yet the neural mechanisms that achieve this feat are poorly understood, and transform-tolerant recognition remains a major stumbling block in the development of artificial vision systems. Even for variations in the position of an image on the retina, arguably the simplest transform that the visual system must discount, little is known about how invariance is achieved.

Several authors have proposed that one solution to the invariance problem is to learn representations through experience with the spatiotemporal statistics of the natural visual world (Foldiak 1991, Wallis & Rolls 1997, Wiskott & Sejnowski 2002, Edelman & Intrator 2003). Visual features that co-vary across short time intervals are, on average, more likely to correspond to different images of the same object than to different objects, and thus one might gradually build up invariant representations by associating patterns of neural activity produced by successive retinal images of an object. While some transformations of an object's retinal image are played out smoothly across time (e.g. scale, pose), changes of an object's retinal position often occur discontinuously as a result of rapid eye movements that sample the visual scene (saccades). A possible strategy, then, for building position-invariant object representations is to associate neural activity patterns across saccades, preferably taking into account the direction and magnitude of the saccade.

If correct position invariance is created through experience with the statistical properties of the

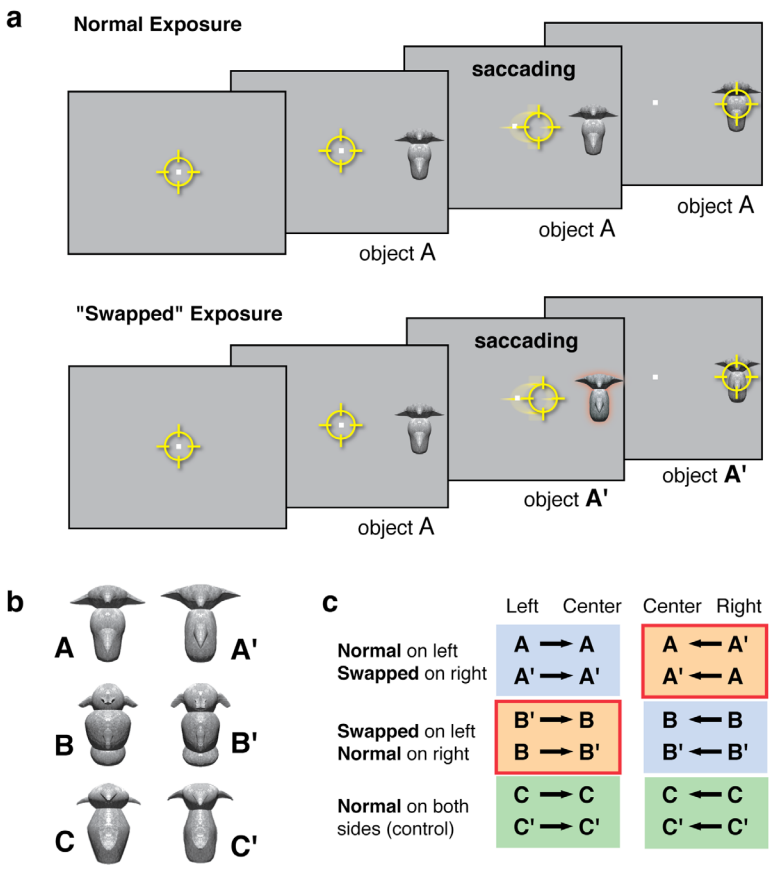


Figure 1. Experiment 1 and 2 design Twelve naïve subjects participated in each experiment and provided informed consent in accordance with the MIT Committee on the Use of Humans as Experimental Subjects. **a)** During the **exposure** phase of each experiment, subjects received two different types of exposure trials randomly interleaved. In all trials, subjects started a trial by fixating on a point, and then an object appeared in the periphery (6° to the left or right, randomly). Subjects spontaneously saccaded to the object, and were required to decide if this object was the same object as in the preceding trial. In **normal exposure** trials, the object identity did not change, so the same object was presented to both the peripheral retina (pre-saccade) and the central retina (post-saccade). In **"swapped" exposure** trials, unknown to subjects, one object was swapped for a different object in mid-saccade, such that one object was presented to the peripheral retina (pre-saccade), and a different object was presented to the central retina (post-saccade). **b)** The objects used in this experiment were modified versions of the publicly available "greeble" stimuli (see Supplemental Methods online) and were arranged in three pairs, with the differences within pair (e.g. A and A') being qualitative smaller than the differences between pairs (e.g. A and B). Objects were chosen to be relatively natural, but unfamiliar to the subject. **c**) A schematic representation of the twelve exposure trial types for one subject. All such exposure trials occurred equally often (pseudo-randomly selected). Thus, each subject received an equal number of presentations of all objects in each retinal location. The letter on one side of the arrow indicates the peripherally presented object (either on the right or left), with the arrow indicating the object identity before (arrow tail) and after (arrow head) the saccade. For all subjects, one object pair was swapped on the right, but normal on the left (first row of boxes), one pair was normal on the right but swapped on the left (second row of boxes), and one pair was not swapped on either side (third row of boxes). Subjects were run in two sets of six, with each set of six counterbalancing across all possible assignments of the three object pairs to each of these three roles.

visual world, it might be possible to create unnatural or "incorrect" invariances by manipulating those statistics. In particular, if objects consistently changed their identity as a function of retinal position, then the visual system might incorrectly associate the neural representations of different objects at different positions into a single object representation. The resulting representation would be activated by one object at one retinal position, and another object at another position, and thus the two objects would be perceived as being the same object at different positions.

In the present study, we engineered such a situation, taking advantage of the fact that humans are effectively blind during the short time it takes to complete a saccade (Ross et al. 2001, McConkie et al. 1996). By monitoring eye position in realtime, we were able to present one object to a subject's peripheral retina that was replaced by a particular different object in mid-saccade when the subject attempted to foveate it. None of the subjects reported being aware that objects were being swapped, despite being asked in a post-session debriefing whether they had seen objects change or appear otherwise unusual. Following a brief period of exposure to these altered spatiotemporal statistics (240-400 altered exposures in Experiment 1, and 120-180 altered exposures in Experiment 2), we used a same-different task to probe the subject's representations of these objects across changes in position. The layout of Experiments 1 and 2 is described in Fig. 1, and in the Supplementary Methods online.

In both Experiments, subjects significantly more often confused object pairs when they were tested across the retinal positions where those particular objects had been swapped during the exposure phase, relative to tests across positions where the same objects had not been swapped (P = 0.0082 in Experiment 1, P = 0.022 in

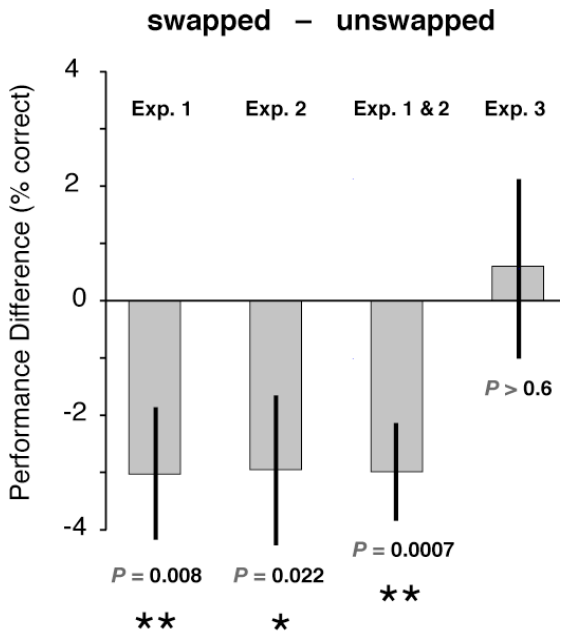
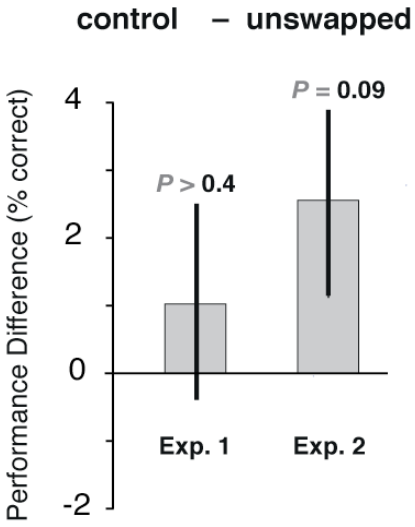


Figure 2. Results In testing following exposure, subjects in Experiment 1 (two days of exposure; n = 12) and Experiment 2 (one day of exposure; n = 12) significantly more often confused objects across retinal positions where they had been swapped during the exposure phase (orange panels in **Fig. 1c**), as compared to the same objects across positions where they behaved normally during exposure ("unswapped"; blue panels in **Fig. 1c**). These effects were not significantly different for "same" trials and "different" trials in either Experiment 1 or 2 (P > 0.4 two-tailed paired t-tests). Subjects in Experiment 3 (replay experiment; n = 12), who received retinal exposure matched to subjects in Experiment 1 did not show a significant effect. Bars show effect magnitudes and standard errors for Experiment 1 (2-day), Experiment 2 (1-day), data from Experiments 1 & 2 pooled together, and Experiment 3 (replay). Mean performance with the control objects was 74%, 72%, and 78%, in Experiments 1, 2 and 3, respectively, and was not significantly different across the three experiments (P > 0.1, one-way ANOVA).

Experiment 2; $P = 0.0007$, both experiments pooled; one-tailed paired t-test; **Fig. 2**). That is, for previously swapped objects, subjects were more likely to perceive different objects at two retinal positions as the same object, and the same object at two positions as different objects.

These results show that confusions in invariant visual object processing occur after relatively brief exposure (< 1 hr total) to altered spatiotemporal statistics across saccades, even though subjects were unaware of this change. Moreover, the confusions are predictable in that they are those expected if the visual system assumes that object identity is stable across the short time interval of a saccade. While the magnitude of the observed effect is not large, and we have only shown it for relatively similar objects, it should be borne in mind that the anomalous exposure provided represents a tiny fraction of each subject's lifetime experience with an unaltered, real-world visual environment. The ability to significantly shift object representations at all suggests that position-invariant visual object recognition is modifiable in adults, and points to possible mechanisms by which sets of invariant features might be acquired, especially during early visual learning.



Supplemental Figure 1. Subjects tended to perform slightly better with object pairs that were never swapped on either side ("control" conditions; green panels in **Fig. 1c**) than with test object pairs across positions where those objects had behaved normally during the exposure phase ("unswapped" conditions; blue panels in **Fig. 1c**), though this trend was not significant in either Experiment. Such a trend suggests at least the possibility that, in addition to effects on position invariance, anomalous exposure may also produce some general deficits with objects (i.e. position-independent effects) or deficits when at least one "misbehaving" position (the fovea in this case) is part of the test.

To test whether the observed effect depends critically on the execution of active eye movements, as opposed to spatiotemporal experience alone, we ran a third set of twelve subjects (Experiment 3) with retinal experience matched to the subjects in Experiment 1, but without saccades. These subjects maintained fixation throughout each trial during the exposure phase, and the retinal positions and timing of object exposure was "replayed," trial-by-trial, from the spatiotemporal retinal experience generated by their counterpart subject in Experiment 1. The testing phase was identical to Experiments 1 and 2. Subjects in Experiment 3 showed no effect of anomalous spatiotemporal experience ($P > 0.6$; 1-tailed paired t-test, **Fig. 2**), suggesting that anomalous experience across saccades may be necessary to produce later confusions in invariant object processing.

Although these results show that specific alterations in object spatiotemporal experience can alter position invariant recognition with test objects in the direction predicted by theory, we wondered if such anomalous experience might also produce

more general deficits in recognition performance with those test objects. To examine this, we compared recognition performance of test objects across positions where those objects had behaved normally ("unswapped" conditions) with recognition of control objects (which were never swapped in either position). Although both experiments showed a trend toward reduced performance with objects whose spatiotemporal statistics had been altered (see Supplemental **Fig. 1**), no significant difference was found in either experiment (Experiment 1: $P = 0.48$; Experiment 2: $P = 0.094$, two-tailed paired t-tests).

Like some recent perceptual learning studies, this study shows that visual processing can be altered by visual statistics that do not reach awareness (Watanabe 2001). However, in contrast to standard perceptual learning paradigms, where subjects improve on some sensory task over the course of many training sessions (Karni & Sagi 1993), here, performance is impaired in a predictable way by brief exposure that runs counter to the subject's past visual experience. This resembles other long-term perceptual adaptation effects, such as the McCollough effect and prism adaptation, and like these effects, might represent an ongoing process to adapt to the environment and keep perception veridical (Bedford 1999).

While adult transform-invariant object recognition is, for the most part, automatic and robust (Biederman and Bar 1999), this finding adds to a growing body of research suggesting that such invariance may ultimately depend upon experience (Dill & Fahle 1998, Nazir & O'Reagan 1990, Dill & Edelman 2001, Wallis & Buelthoff 2001). More broadly, this finding supports the developing belief that visual representations in the brain are plastic and largely a product of the visual environment (Simoncelli & Olshausen 2001). Within this context, invariant object representations are not rigid and finalized, but are continually evolving entities, ready to adapt to changes in the environment.

Methods

Subjects. Twelve naïve subjects participated in each of the three experiments (ages 18-45; normal or corrected-to-normal vision, 9 male, 27 female), and provided informed consent in accordance with the MIT Committee on the Use of Humans as Experimental Subjects.

Stimuli. The objects used in this experiment were modified versions of the publicly available "greeble" stimuli[1]. Object images were approximately 2.5º wide by 4º high and were presented on a gray background on a 21" Trinitron CRT (viewing distance of approximately 65 cm), using custom software that also handled saccade detection, swapping of stimuli in mid-saccade, response collection and all other aspects of the experiment. The same three pairs of objects were used in all three experiments, with subtle differences between members of each pair, and qualitatively greater differences between pairs (**Fig. 1b**). Relatively similar objects were used for each pair under the logic it might not be possible to induce radical shifts in object representations within the time constraints of an experimental session.

Eye Tracking. Subjects' eye positions were tracked using an EyeLink II head-mounted infrared video eye tracking system (SR Research Ltd., Mississauga, ON, Canada) running at a rate of

250 Hz, with built in head-tracking. Subjects were required to fixate within 1.5º of the central fixation point, and trials were aborted if the subject's eye position deviated from this window. The endpoints of saccades to objects were repeatedly estimated in mid-saccade, and trials were aborted (i.e. the object was removed) on trials where saccades were estimated to be headed to land outside of the target object.

Exposure Phase (Experiments 1 & 2). During the exposure phase, an object appeared 6º to the left or right of the fixation point (randomly, see **Fig. 1a**). Subjects had been instructed to feel free to look at any object and, to ensure that they attended to the objects, decide if it was the same object that had appeared in the previous trial. Unknown to subjects, some objects were replaced by the other member of their pair while the subject was making this saccade (see **Fig. 1a**). Thus, Object A might appear to the left of fixation, eliciting a saccade to the object, but be replaced by Object A' by the time the subject's eyes landed. Each subject experienced i) one of the three pairs of objects swapped in mid-saccade on the left, but behaving normally on the right, ii) another pair swapped on the right but not on the left, and iii) the third pair not swapped in either position (control pair, see **Fig. 1c**). Each subject experienced each of the six objects equally often in each position (see **Fig. 1c**), and object pairs were counterbalanced across subjects such that each of the three object pairs was equally often swapped on the left, swapped on the right, or not swapped at all. None of the subjects reported being aware that objects were being swapped, despite being asked in a post-session debriefing whether they had seen objects change or appear otherwise unusual. Subjects did not take longer to saccade to the to-be-swapped objects ($P > 0.4$, mean: 200.7 ms), nor did they did look at swapped objects for a significantly different amount of time ($P > 0.8$, mean: 354.5 ms).

"Replay" Exposure Phase (Experiment 3). Twelve subjects in Experiment 3 were each paired with one of the twelve subjects in Experiment 1 and received retinal exposure that was matched, trial for trial, to their counterpart in Experiment 1. Subjects in this experiment were instructed to fixate the central fixation point while objects appeared first in the periphery, and then at the center of gaze with timing generated from the saccades made by their counterpart subject in Experiment 1. The screen was left blank during the time that the Experiment 1 subjects' eyes had been in flight, simulating the lack of appreciable form vision while the eyes are moving at high velocity. Failures to maintain fixation resulted in the trial being aborted and re-run. Subjects performed an analogous 1-back task as in Experiment 1, in which they reported whether the object was the same or different than the object presented on the previous trial. The instructions implied that the same object would appear in the periphery and at the center of gaze, even though different objects would in fact appear on the "swapped" trials.

Testing Phase (Experiments 1, 2 & 3). During the testing phase, designed to probe object representations across retinal positions, subjects fixated while an object appeared briefly in the periphery (6º; 150 ms), followed by a 300 ms delay, and then either the same object or the other member of its pair at the center of gaze (150 ms). Subjects indicated whether the two objects were the same or different. No feedback was given regarding accuracy of their responses. Each testing block contained an equal number of all combinations of within-object-pair comparisons, and peripheral positions (right and left). Blocks where subjects did not perform significantly above chance with control objects ($P > 0.2$; less than 5% of all data) were excluded from further

analysis.

Experimental Sessions. Experiment 1 was conducted across two days, with subjects receiving exposure on both days (720-1200 total exposure trials, of which 240-400 were swapped) and completing four testing blocks (120 trials each) at the end of the second day. Experiment 2 was conducted in a single session, with subjects receiving 360-540 exposure trials (120-180 swapped exposures), and completing three testing blocks (120 trials each) at the end of the same session. In Experiment 3, the number of sessions, training blocks, and testing blocks was exactly matched to Experiment 1.

Chapter 3:
Multiple Object Response Normalization in Monkey Inferotemporal Cortex

This chapter originally appeared as:

Zoccolan DZ, Cox DD*(contributed equally), DiCarlo JJ (2005). "Multiple object response normalization in monkey inferotemporal cortex" Journal of Neuroscience 25(36):8150–8164*

Copyright 2005 by the Society for Neuroscience.

The highest stages of the visual ventral pathway are commonly assumed to provide robust representation of object identity by disregarding confounding factors such as object position, size, illumination and the presence of other background objects (clutter). While robust tolerance to position and size changes has been reported for neuronal responses in the monkey inferotemporal cortex (IT), previous studies report that IT neuronal responses to preferred objects are usually weaker when a non-preferred object is present. However, we lack a systematic understanding of multiple object representation in IT and it is not known how to explain IT responses to multiple objects based on responses to those same objects in isolation. In this study, we systematically examined IT neuronal responses to the presentation of pairs and triplets of objects in three passively viewing monkeys across a broad range of stimulus conditions. Our results show that a large fraction of IT neuronal responses to multiple objects can be reliably predicted as the average of the responses to the constituent objects in isolation. That is, each IT neuron's response to multiple objects depends largely on the relative effectiveness of each of the constituent objects, and it does not matter if that effectiveness is altered by changing object shape or the RF position at which an object is presented. These observations are consistent with mechanistic models in which the output of each IT neuron is normalized by a summation of synaptic drive into IT or spiking activity in IT and suggest that normalization mechanisms previously revealed at earlier visual areas may be operating throughout the ventral visual stream.

Introduction

Visual object recognition in cluttered, real-world scenes remains an extremely difficult problem for artificial vision systems, yet is somehow effortlessly solved by the brain. In primates, it is believed that object identity is extracted through processing along the ventral visual stream and that it is explicitly represented in patterns of neuronal activity in the highest stages of that stream

– the anterior inferotemporal cortex (IT). This hypothesis is based on IT lesion studies (Dean, 1976, 1982; Weiskrantz and Saunders, 1984; Horel, 1996) showing impaired visual recognition and neurophysiological studies showing that IT neurons can be highly selective for complex objects while also being largely tolerant to some transformations (object position, scale, and pose; for review, see Logothetis et al., 1994; Logothetis et al., 1995; Logothetis and Sheinberg, 1996; Tanaka, 1996). In this context, it has been suggested that idealized IT neurons should also be tolerant to visual clutter (Rousselet et al., 2003, 2004). That is, if each IT neuron participates in the representation of some object or subset of objects, then, ideally, its response to that object(s) should be largely unaffected by the presence of other objects. However, this idealized notion of IT does not appear to be correct in that IT responses are altered in cluttered scenes (Sheinberg and Logothetis, 2001; Rolls et al., 2003), and responses to pairs of simultaneously presented objects are typically weaker than responses to the preferred object presented alone (Sato, 1989; Miller et al., 1993; Rolls and Tovee, 1995; Missal et al., 1997; Chelazzi et al., 1998; Missal et al., 1999).

Despite the obvious relevance of IT clutter-tolerance properties to theories of object representation, we do not have a systematic understanding of these properties, even in clutter conditions that involve only two objects. For example, one study reported that response suppression caused by the addition of a second object in the RF did not depend on the shape of that object (Miller et al., 1993), while another study reported the opposite (Missal et al., 1999). Similarly, while some studies hint at a systematic relationship between the response to object pairs and the response to the constituent objects (Miller et al., 1993; Rolls and Tovee, 1995; Chelazzi et al., 1998), another study (Missal et al., 1999) explicitly ruled out that possibility. Moreover, no study has systematically tested this relationship over a full range of stimulus effectiveness or with a large battery of testing stimuli, and there has been no attempt to understand how IT responses to multiple objects depend on the shape similarity of those objects. Although progress on understanding responses to multiple stimuli has been made in area V4 (Reynolds et al., 1999; Reynolds and Desimone, 2003), a recent study calls those results into question by suggesting that, like idealized IT neurons, V4 responses to preferred stimuli are largely tolerant to the presence of an additional stimulus (Gawne and Martin, 2002).

Although other studies have increased our understanding of how visual attention modulates processing of targets in the presence of distractors (Moran and Desimone, 1985; Desimone and Duncan, 1995; Maunsell, 1995; Connor et al., 1997; Chelazzi et al., 1998), such work has not provided a systematic characterization of neuronal responses to multiple objects (but see Reynolds et al., 1999; Reynolds and Desimone, 2003). Notably, visual recognition shows remarkable clutter tolerance even for brief presentation conditions (e.g. ~100 ms) without explicit attentional instruction (e.g. Potter, 1976; Intraub, 1980; Rubin and Turano, 1992). This strongly suggests that, besides top-down attentional mechanisms, powerful, largely feedforward, clutter-tolerance mechanisms are at work. An understanding of these "core" mechanisms – the rapid IT population response in clutter – is not only fundamental, but should lead to improved understanding in situations where attention is manipulated.

In this study, we systematically examined the IT neuronal responses to rapid presentation of multiple objects in three passively viewing monkeys using two complementary experimental para-

digms. Our results show that, across a wide range of stimulus conditions, IT neuronal responses to multiple objects are very well predicted by the average of their responses to the constituent objects. Moreover, most IT neurons are not idealized object detectors in that they do not respond equally to preferred objects in spite of the presence of non-preferred objects. These observations suggest that divisive normalization mechanisms analogous to those proposed to explain response re-scaling in early visual stages (Heeger, 1992; Desimone and Duncan, 1995; Heeger et al., 1996; Carandini et al., 1997; Reynolds et al., 1999; Schwartz and Simoncelli, 2001; Cavanaugh et al., 2002) and area MT (Recanzone et al., 1997; Britten and Heuer, 1999; Heuer and Britten, 2002) could operate in IT.

Methods

Animals and surgery.

Experiments were performed on three male rhesus monkeys (*Macaca mulatta*) weighing approximately 8, 9.5 and 10kg. Before behavioral training, aseptic surgery was performed to attach a head post to the skull of each monkey and to implant a scleral search coil in the right eye of monkeys 1 and 2. After 2-5 months of behavioral training (below), a second surgery was performed to place a recording chamber (18 mm diameter) to reach the anterior half of the left temporal lobe (chamber Horsley-Clark center = 15 mm A). All animal procedures were performed in accord with National Institute of Health guidelines and the Massachusetts Institute of Technology Committee on Animal Care.

Eye position monitoring

Horizontal and vertical eye positions were monitored using the scleral search coil (monkeys 1 and 2) or a 250 Hz camera based system (monkey 3; EyeLink II, SR Research Ltd., Osgode, ON, Canada). Each channel was digitally sampled at 1 kHz. Methods for detecting saccades and calibrating retinal locations with monitor locations are described in detail elsewhere (DiCarlo and Maunsell, 2000).

Visual stimuli.

Stimuli were presented on a video monitor (43.2 x 30.5 cm, 75 Hz frame rate, 1920 x 1200 pixels) positioned at 81 cm from the monkeys (so that the display subtended approximately ±15 (h) and ±10 (v) deg of visual angle). Different visual objects were used in each experiment (see Fig.1 and below).

Experiment 1. Monkeys 1 and 2 were tested with three simple, solid geometric forms (a star, a cross, and a triangle, see Fig. 1A, left), presented at full-luminance (57 Cd/m^2) on a gray background (27 Cd/m^2). Each object was 2º in size (diameter of a bounding circle).

Experiment 2. Monkey 3 was tested using objects drawn from three object sets with parametrically controllable shape similarity within each set (Fig. 1A, right). To assure generality of results, three different spaces of morphed shapes were generated: 1) a *car* space; 2) a *face* space;

3) a *NURBS* space (non-uniform rational B-spline generated two-dimensional silhouettes). Each space was generated from a set of 15 initial shapes: 1) 15 three-dimensional models of car brand prototypes; 2) 14 three-dimensional models of human heads plus their average; 3) 15 randomly generated NURBS (44 free parameters, see below). For each space, one of these initial shapes

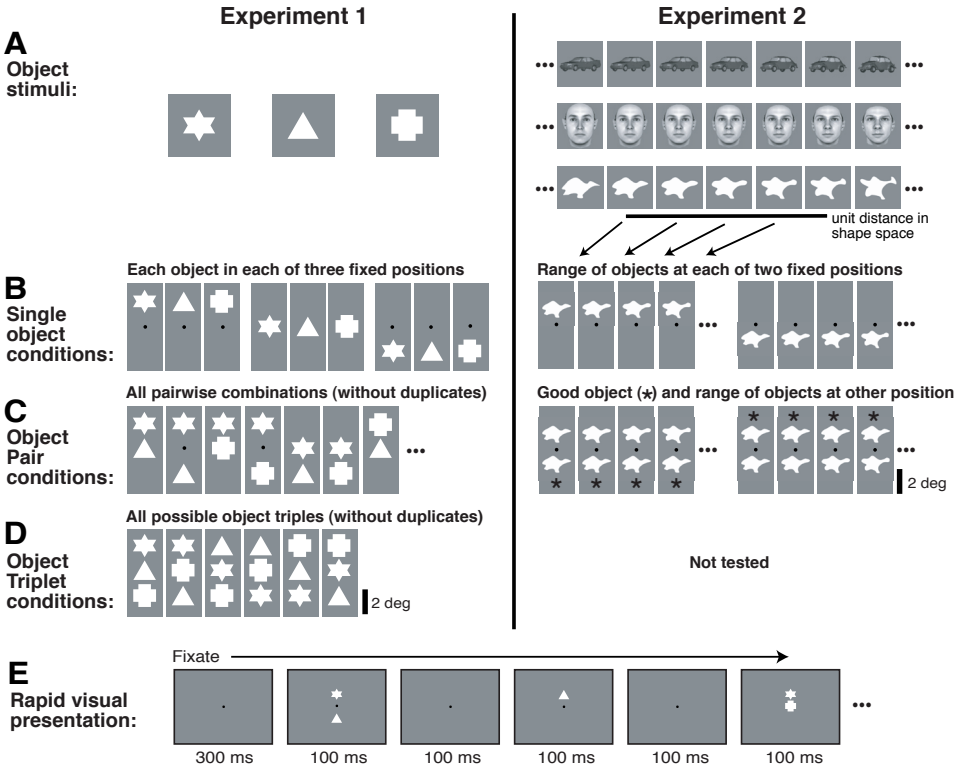


Figure 1. Stimulus conditions during recordings. A. Left. The three geometrical shapes used in Experiment 1. Right. An example morph-line (i.e., set of parametric shapes) from each of the three shape spaces (i.e., cars, faces and 2D silhouettes) used in Experiment 2. The horizontal line indicates the unit shape distance within a morph-line. This is the distance between the object prototypes used to generate the morph-line (i.e., the 2nd and 6th stimulus in each row). B. Single object conditions. Left. All 9 single object arrangements of Experiment 1 (3 shapes in each of 3 visual field locations: at center of gaze and 2º above and below center of gaze). Right. Single objects sampled from the most selective morph-line (in the example, 2D silhouettes) were presented in two visual field locations: 1.25º above center of gaze (top) and 1.25º below center of gaze (bottom) in Experiment 2. C. Object pair conditions. Left. A subset of the 18 object pairs used in Experiment 1 (3 objects in 2 of 3 positions without duplicate objects). Right. Examples of objects pairs used in Experiment 2. In each pair, the neuron's "preferred" object (indicated by the asterisk) is presented in either the top or bottom position and is paired to a second object drawn from a range of shapes along the morph-line containing the "preferred" object. D. Object triplet conditions. Left. All 6 object triplet arrangements used in Experiment 1 (the 3 objects in the 3 positions without duplicate objects). Right. No object triplets were tested in Experiment 2. E. Rapid visual presentation. Each panel is a schematic of the visual display (not to scale). The monkey was required to hold fixation on a central point while stimulus conditions were randomly interleaved and presented at a rate of 5 per second (see Methods).

was chosen as "center" of the space and 14 sets of morphed shapes were built as blends (see below) of the center shape and each of the other 14 prototype shapes, thus resulting in 14 morph-lines per space (see examples in Fig. 1A, right). In each of the three object spaces, the distance between the center shape and each of the 14 prototype shapes was defined to have value 1. As shown for the three exemplar morph-lines of Figure 1A, morphed shapes were generated not only between the "center" and each of the 14 prototypes (e.g. the five middle shapes in each row in Fig. 1A, right) but also by extrapolating beyond the initial prototypes (first and last shapes in each row in Fig. 1A, right), thus resulting in shape distances $d > 1$ and $d < 0$.

Slightly different morphing methods were used to generate the objects in each of the three shape spaces. Cars were built using an algorithm (Shelton, 2000) that found corresponding points in each pair of 3D car prototypes and represented each car prototype as a vector of point co-ordinates. Car morphs were created as linear combinations of these vectors, then rendered as grayscale 2D images (with fixed viewpoint, illumination and size; first row in Fig. 1A, right). Faces were generated by a face morphing algorithm (Blanz and Vetter, 1999), in which point correspondences between pairs of face prototypes were established based on the three-dimensional structure of the head models. Face morphs were created as linear combinations of corresponding points in the head pairs, then rendered as grayscale 2D images (with fixed viewpoint, illumination and size; second row in Fig. 1A, right). The center shape of the face space was the average face (second stimulus in the second row of Fig. 1A, right). NURBS objects were filled shapes defined by closed third-order NURBS curves with 22 equally-weighted control vertices (Rogers, 2000). NURBS morphs were generated using weighted averages of control vertices of pairs of prototypes and all NURBS curves were filled at full luminance (72 Cd/m^2; third row in Fig. 1A, right). All objects were presented at 2º in size (bounding circle diameter) on a gray background (12 Cd/m^2).

Behavioral task and training

All three monkeys were trained to fixate a central point (0.2 x 0.2 deg), for several seconds while a series of visual stimuli were presented in rapid succession (rapid, passive viewing paradigm). In particular, stimulus conditions were presented in a random sequence where each stimulus condition was on for 100 ms, followed by 100 ms of a gray screen (no stimulus), followed by another stimulus conditions for 100 ms, etc. (see Fig. 1E). That is, stimulus conditions were presented at a rate of 5 per second. At this presentation rate, IT neurons show robust object selectivity (Keysers et al., 2001) and this rate is consistent with that produced spontaneously by free viewing monkey (DiCarlo and Maunsell, 2000). Single, pair and triplet object conditions were pseudoranomly interleaved (see schematic in Fig. 1E). The screen background was always kept at a constant gray. The total number of stimulus conditions presented on each fixation trial ranged from 3 to 20 and the monkey was rewarded for maintaining fixation throughout the trial (±0.5º fixation window in Monkeys 1 and 2; ±1.5º fixation window in Monkey 3). Failures to maintain fixation throughout the trial resulted in the trial being aborted, and all stimulus conditions in that trial were re-presented.

The data presented in the current study were all acquired during this rapid, passive viewing paradigm. However, all three monkeys are also involved in ongoing studies that require behavioral training with the stimuli used in this study. Monkeys 1 and 2 were trained to perform an object

identification task with single geometrical shapes presented either at the center of gaze, 2º above, or 2º below fixation. Monkeys were required to saccade to a different, fixed peripheral target for each object. Monkey 3 was trained to perform a sequential object recognition task that required the detection of a fixed target shape (the "center" object in each object set) embedded in a temporal sequence of shapes drawn from the same object set (blocked trials).

Recording and data collection

For each recording, a guide tube (26 G) was used to reach IT using a dorsal to ventral approach. Recordings were made using glass-coated Pt/Ir electrodes (0.5-1.5 MΩ at 1 kHz) and spikes from individual neurons were amplified, filtered, and isolated using conventional equipment. The superior temporal sulcus (STS) and the ventral surface were identified by comparing gray and white matter transitions and the depth of the skull base with structural MR images from the same monkeys. Penetrations were made over a ~10x10 area of the ventral STS and ventral surface (Horsley-Clark AP: 10-20 mm, ML: 14-24 mm) of the left hemisphere of each animal. All recordings were lateral of the Anterior Middle Temporal Sulcus (AMTS). Thus, the recorded regions included AIT and CIT (Felleman and Van Essen, 1991). In all three animals, the penetrations were concentrated near the center of this region, where form selective neurons were more reliably found. The animals cycled through behavioral blocks as the electrode was advanced into IT. Responses from every isolated neuron were assessed with an audio monitor and online histograms, and data were collected according to specific criteria for Experiment 1 and 2.

Experiment 1. As the electrode was advanced into IT, Monkeys 1 and 2 performed the object identification task described above. Neurons that responded to any of the geometric objects at any of the three positions were further probed while the animal passively viewed the same objects (described above; see Fig. 1E). Neurons that responded with mean firing rate significantly higher than background rate to any shape at any position (t-test, $p < 0.05$) were studied further. The main experimental conditions included the following: 1) each of the three shapes presented in isolation in each of three positions (Fig. 1B, left; 3 shapes x 3 positions = 9 stimulus conditions); 2) pairs of objects in all possible arrangements that did not include object duplicates (Fig. 1C, left; 18 stimulus conditions); and 3) triplets of objects in all possible arrangements that did not include object duplicates (Fig. 1D, left; 6 conditions). Object size (2 deg) and positions (fixation, 2º above fixation and 2º below fixation) were chosen before data collection so that the objects did not touch or overlap but that objects were close enough to likely activate IT neurons in one or more positions. No attempt was made to optimize the objects or positions for the neuron under study. Instead, the exact same 33 stimulus conditions were tested for each neuron. These conditions were pseudo-randomly interleaved and presented using the rapid, passive viewing paradigm described above. All neurons in which these conditions were tested were considered in the Results if 10-30 presentations of each condition were completed during the time that the neuron was isolated.

Experiment 2. As the electrode was advanced into IT, Monkey 3 was either engaged in the rapid, passive viewing paradigm or engaged in a recognition task similar to the behavioral task described above (except that the target object was a red triangle). To search for neurons with strongly selective responses across at least one of the morph-lines, each isolated neuron was tested with a sequence of screening procedures that always included at least 10 repetitions of each

stimulus condition (pseudorandomly interleaved). During the first screening, 15 objects from each morphed space (a total of 45 objects) were presented at the center of gaze. These 15 objects were the center shape (see above) plus one stimulus randomly sampled (at a distance of 0.5 or 1.0 from the center object) from each of the 14 morph-lines. Neurons who responded to one of these stimuli with mean firing rate significantly higher than background rate (t-test, $p < 0.005$) were further tested using objects within the space to which the most effective stimulus belonged (all tested during the rapid, passive viewing paradigm described above). In particular, the center object and 4 objects sampled (at distances $d = 0.25, 0.5, 0.75, 1$ from the center object) from each of the 14 morph-lines were presented in isolation at the center of gaze. A neuron was considered to be selective if the mean firing rates elicited by the set of 5 objects belonging to at least one of the morph-lines were significantly different (ANOVA, $p < 0.05$). If so, the object along this morph-line that was most effective in driving the cell was taken to be the neuron's "preferred object" and more tests of object selectivity were done using objects drawn from this morph-line.

The main experimental conditions in Experiment 2 included the following two primary conditions: 1) 8 – 12 isolated objects from the most selective morph-line (morphing step distance d_{step} ranging from 0.1 to 0.5). For most neurons, this set of objects included shapes generated by moving beyond the limits of the initial morph-line, as well as one randomly chosen object from one of the two other object sets. Each object was presented at each of two, fixed positions (1.25 deg above the center of gaze and 1.25 deg below the center of gaze; Fig. 1B, right). Thus, a total of 16 – 24 isolated object conditions were tested for each neuron. As in Experiment 1, object size (2 deg) and positions were chosen and fixed before data collection so that the objects did not touch or overlap but that objects were close enough to likely activate IT neurons in one or more positions. However, unlike Experiment 1, the tested range of objects was both parameterized (morph-line) and chosen to obtain maximal selectivity from each neuron. 2) Pairs of objects were presented to all neurons to systematically test each neuron's ability to tolerate the presence of a second object given the presence of a "preferred" object. In particular, the neuron's "preferred object" (resulting from the previous screening at fovea) was presented at one position in combination with each of the objects tested in isolation (see above), including the preferred object itself (8 – 12 conditions; see Fig. 1C, right). This was also done with the preferred object in the other position (Fig. 1C, right). In sum, a total of 16 – 24 isolated object conditions and 16 – 24 paired object conditions were tested for each neuron. 15-30 repetitions of each stimulus condition were recorded for each neuron (pseudorandomly interleved) using the rapid, passive viewing paradigm described above.

Analysis

Only neuronal responses collected during correctly completed behavioral trials were included in the analysis. The background firing rate of each neuron was estimated as the mean rate of firing over all trials in a 100 ms duration window that directly preceded the onset of the first stimulus in each trial. For all the data recorded from the three monkeys, we quantified the response of each neuron to each of the stimulus conditions as the mean firing rate in a 100 ms window that began 100 ms after stimulus onset. The statistical tests used to assess neuronal responsiveness and selectivity to the different stimulus conditions are explained in the Results, as well as the criteria to include subsets of recorded neurons in each analysis. In the following, details about some of the analysis carried out in the Results are provided.

Goodness Of Fit analysis (GOF). To assess, for each neuron, how much of the variance of the responses to objects pairs could be accounted by considering responses to the constituent objects presented in isolation, a Goodness Of Fit (GOF) index was computed. The GOF index provides an unbiased estimate of the percentage of true data variance explained by a given model, by removing the fraction of data variance that is merely due to noise (i.e., the trial-by-trial variability of the neuronal response). The GOF index calculation is based on well known mathematical relationships that are at the base of the ANOVA statistics. Following the convention used by (Rice, 1995), let us assume we recorded J neuronal responses to each of I different stimulus pairs (I and J are, respectively, the number of groups and trials in the ANOVA statistics). Let σ^2_{expl} be the true (or "explainable") variance of the mean recorded responses to the stimulus pairs. Let σ^2_{noise} be the variance of the noise that contaminates neuronal responses. Let SS_B and SS_W be the sum of squares, respectively, between groups and within groups of the ANOVA statistics for the recorded responses. The following relationship holds for the expectation of SS_B: $E(SS_B) = J(I\text{-}1)\ \sigma^2_{expl} + (I\text{-}1)\ \sigma^2_{noise}$ (Rice, 1995). Since the noise variance can be estimated as $\sigma^2_{noise} = SS_W / [I(J\text{-}1)]$, the explainable variance can be estimated as: $\sigma^2_{expl} = SS_B / [J(I\text{-}1)] - SS_W / [IJ(J\text{-}1)]$.

Given a model providing a prediction for the mean response to each object pair, the deviations from the model predictions can be computed for each trial J and each group (stimulus pair) I, so as to obtain trial-by-trial residual responses to each stimulus pair. Again, the variance of the mean residual responses to the stimulus pairs is composed of two terms: the noise variance σ^2_{noise} and the variance σ^2_{res} of the true deviations from the tested model. Therefore, σ^2_{res} can be estimated by the same equation that gives σ^2_{expl}, but with SS_B and SS_W obtained for the ANOVA statistics of the residual responses.

Once σ^2_{res} and σ^2_{expl} are estimated from the data, the GOF index can be computed as: $GOF = 100\ [1\text{-}\ \sigma^2_{res} / \sigma^2_{expl}]$. We verified that this method provides an unbiased estimate of the percentage of explainable variance explained by a model, by running simulations in which data points were generated according to a linear model contaminated by different amounts of noise.

The standard error (SE) of the GOF index was estimated by bootstrap resampling. For each of the I stimulus pair conditions, J responses were re-sampled with replacement 200 times from the J responses obtained during recordings. The GOF index was computed for each of these re-drawing of the response matrix and the standard deviation of the resulting 200 bootstrapped GOF indexes was taken to be the SE of the GOF (Efron and Tibshirani, 1998).

Selectivity and monotonicity criteria for the tuning curves included in the population averages. Neurons recorded in Experiment 2 were tested with parametric objects sampled from morphed object spaces (see above). Therefore, tuning curves of neuronal responses to objects along continuous, parameterized changed in object shape (i.e. along a morph-line) were obtained. The range of shape distances spanned by each morph-line during the probing phase of the recordings in the parafoveal positions varied from neuron to neuron. However, each morph-line spanned at least a unit shape distance (horizontal line in Fig. 1A, right) and included the neuron's preferred stimulus obtained from the screening phase of the recordings, whose shape distance was defined as $d = 0$ (see above). To get a meaningful population average of the neuronal tuning properties, the following criteria were used to include each tuning curve in the final average: 1) Responses across all tested single object conditions (i.e., both top and bottom positions) were highly selec-

tive (ANOVA, $p < 0.001$). 2) The tuning curve in the tested position was significantly selective in a shape range spanning the unit distance (i.e., in $d \in [0, 1]$; ANOVA, $p < 0.05$). 3) The tuning curve was approximately monotonic in $d \in [0, 1]$, with peak at or near the "preferred" stimulus (i.e., at $d <= 0.25$).

Simulated neuronal responses. One goal of the present study was to understand if neuronal responses to pairs of objects could be more reliably modeled as: 1) the *average* of the responses to the constituent objects presented in isolation (*average* model), or 2) the *maximum* of the responses to the constituent objects presented in isolation (Complete Clutter Invariance model – CCI model). To understand how well measures of explained variance or transformations of the data were suitable for comparing these two models, we simulated neuronal responses to object pairs that followed either the average model or the CCI model (see Fig. 6B). The response of each model neuron to single objects was assumed to have some tuning across a hypothetical continuous shape dimension (a Gaussian tuning was assumed, but any arbitrary tuning function could be used). Then, the response R_{AB} to each pair of stimuli A and B sampled from the same shape dimension was modeled as: 1) the linear sum of individual responses, i.e. $R_{AB} = p + m\ (R_A + R_B)$, or 2) the maximum of individual responses, i.e. $R_{AB} = \mathrm{MAX}(R_A, R_B)$. Random fluctuations (zero mean noise) were added to the responses of the model neurons to the pairs, to simulate more realistic neuronal responses.

Results

Complete recordings using our battery of visual conditions were performed from 104 well-isolated single IT neurons of three monkeys (35 cells in monkey 1, 33 cells in monkey 2 and 36 cells in monkey 3). During recordings, all neurons were tested with both single and multiple objects using rapid visual presentation according to one of the two experimental paradigms (see Fig. 1 and Methods). Each recorded neuron was tested for responsiveness to single objects and neurons that responded significantly to at least one of the presented single objects (relative to background rate) were included in the analyses described through the paper (t-test on each single object condition, $p < 0.05$; 79 of 104 neurons; 29 of 35 cells in monkey 1, 19 of 33 neurons in monkey 2, and 31 of 36 neurons in monkey 3). This weak inclusion criterion without correction for multiple tests was done to minimize sampling bias in that even neurons with weak responsivity were considered.

Responses to pairs of objects

In Experiment 1, no attempt was made to optimize the objects or retinal positions for each neuron. Instead, the same three objects (see Fig. 1A, left) were presented in each of three fixed positions to each neuron (center of gaze and 2 deg above and below the center of gaze). Using those same objects and retinal positions, all pair wise and triple wise combinations were also tested (see Methods and Figs. 1B-D, left). That is, a total of 33 stimulus conditions were tested for all isolated neurons (9 single object conditions, 18 object pair conditions, and 6 triple object conditions). Figures 2A-B shows the response of a typical IT neuron to some of these conditions. For this neuron, the single object that produced the strongest response was the *cross* located at the

center of the gaze (middle panel in Fig. 2A). When the cross was flanked by a non-preferred object located in one of the eccentric positions (2º above or below fixation; first and third panel in Fig. 2B), the response to the resulting object pair was intermediate between the responses to the individual constituent shapes. Similar intermediate responses were observed when the cross was flanked by two non-preferred objects (last panel in Fig. 2B).

Intermediate responses to multiple objects (relative to the responses to single objects) were also obtained in Experiment 2 using sets of objects with parametrically-defined shape similarity (see examples in Fig. 1A, right) that were presented in isolation or in pairs at two fixed retinal positions (see Methods and Figs. 1B-C, right). Like Experiment 1, the same retinal positions were tested for all neurons, but, in contrast to Experiment 1, the presented objects were optimized for the neuron under study. Specifically, a large range of objects was tested across three object spaces

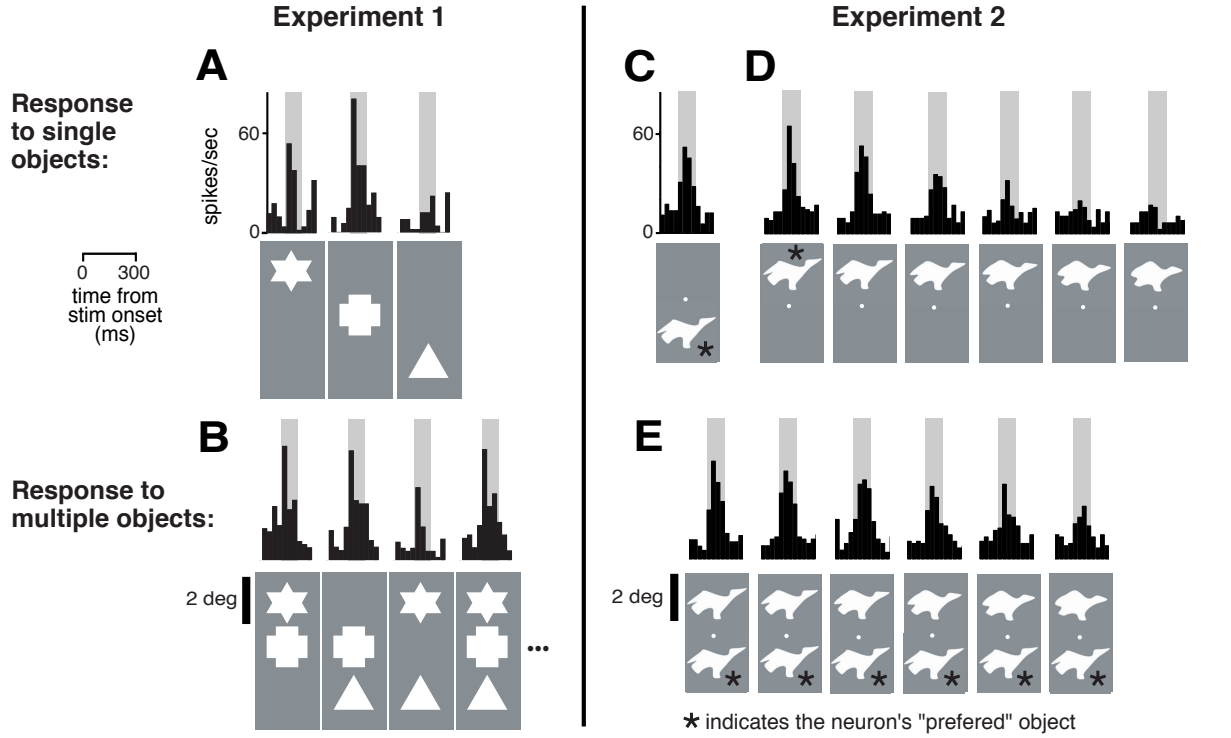


Figure 2. Examples of IT neuronal responses to single and multiple objects. A. The black histograms are the average firing rates (computed in time bins of 25 ms) of a neuron recorded in Experiment 1, following presentation of some of the single object conditions (stimuli are shown below the histograms). Objects were presented at time 0 and the neuron's average response was computed between 100 and 200 ms (gray patch). B. Examples of responses of the same neuron to object pairs and triplets. C. Response of a neuron recorded in Experiment 2 to its "preferred" object presented in the bottom position. D. Responses of the same neuron to a range of objects sampled from the morph-line containing the preferred object and presented in the top location. The neuron's response decays as the second object is made more dissimilar to the "preferred" object (indicated by the asterisk). E. Responses of the same neuron to stimulus pairs composed by the preferred object (asterisk; bottom position) and the range of shapes previously shown in D (top position). In both B and E, responses to the object pairs are intermediate between responses to the constituent objects of the pairs.

(see Methods and Fig. 1A, right) and the set of objects (morph-line) that yielded the most reliable selectivity was studied in detail. Figures 2C-E show a typical activation pattern of an IT neuron recorded in Experiment 2. The response of this neuron to individual objects was significantly selective (ANOVA, $p<0.01$) across a set of eleven objects sampled at consecutive distances along one of the morph-lines of the NURBS space (responses to six of the eleven stimuli are shown in Fig. 2D). The selectivity pattern was unchanged and significant in all three tested locations (centre of the gaze, 1.25º above the center of gaze (top); and 1.25º below the center of gaze (bottom); data not shown). The neuron responded maximally to objects at one extreme of the shape space (the "preferred" shape; Fig. 2C and first histogram in Fig. 2D), while the response to the other extreme was not significantly higher than background (last histogram in Fig. 2D; t-test, $p>0.05$). Responses between these two extremes of object shape showed an approximately monotonic decrease from maximal response as the object was made more dissimilar to the preferred shape (Fig. 2D). The neuron's response to pairs of objects was tested by presenting stimuli containing both the preferred object (bottom position) together with a non-preferred object (top position) sampled across the whole morph-line. The resulting activation pattern is shown in Figure 2E. For each stimulus pair, the neuron's response was intermediate between its responses to the individual constituent shapes of the pair (compare Fig. 2E and 2D). A nearly identical response pattern was obtained when the identity of the object in the bottom position was varied while the preferred object was presented in the top position (data not shown).

To determine if a systematic relationship existed between responses to individual objects and multiple objects, we first plotted the response to each object pair against the sum of the responses to the constituent objects of the pair. Figures 3A and B (first panel) show the resulting scatter plots for the two neurons just described, respectively, in the left and right side of Figure 2. As expected from previous studies, responses to object pairs were smaller than the simple sum of individual responses (i.e. well below the diagonal dashed lines in Fig. 3). Nevertheless, the responses to each object pair condition (18 conditions and 22 conditions in these two cases) did not fall haphazardly on the scatter plot, but clustered along a line of slope 0.5 (solid line). That is, the response of these neurons to pairs of objects was in good agreement with the average of the responses to the constituent objects presented in isolation. To examine this more closely, we considered object pair conditions where each of the two constituent objects drove the neuron significantly above background when presented alone (red dots) and conditions in which only one of the two objects did (blue dots). This did not reveal any obvious difference between such conditions in that both sets of points cluster along the same line. Moreover, the fact that the blue points are well below the diagonal shows that objects that have no significant effect on IT neuronal responses when presented alone can strongly impact responses to more preferred objects. Similar relationships were obtained for most of the neurons recorded in the three monkeys. Figure 3 also shows two additional examples of data from single IT neurons, whose responses to objects pairs were again in good agreement with the average of the responses to the constituent objects over a range of conditions.

As a first look at our entire population of IT neurons in the three monkeys, we pooled the data from all 79 responsive neurons in a scatter plot using the same axes shown for the example neurons (Fig. 3C). Like the individual examples, responses to pairs of objects were highly correlated with the sum of responses to the constituent objects ($r = 0.92$) and the slope of the best linear fit

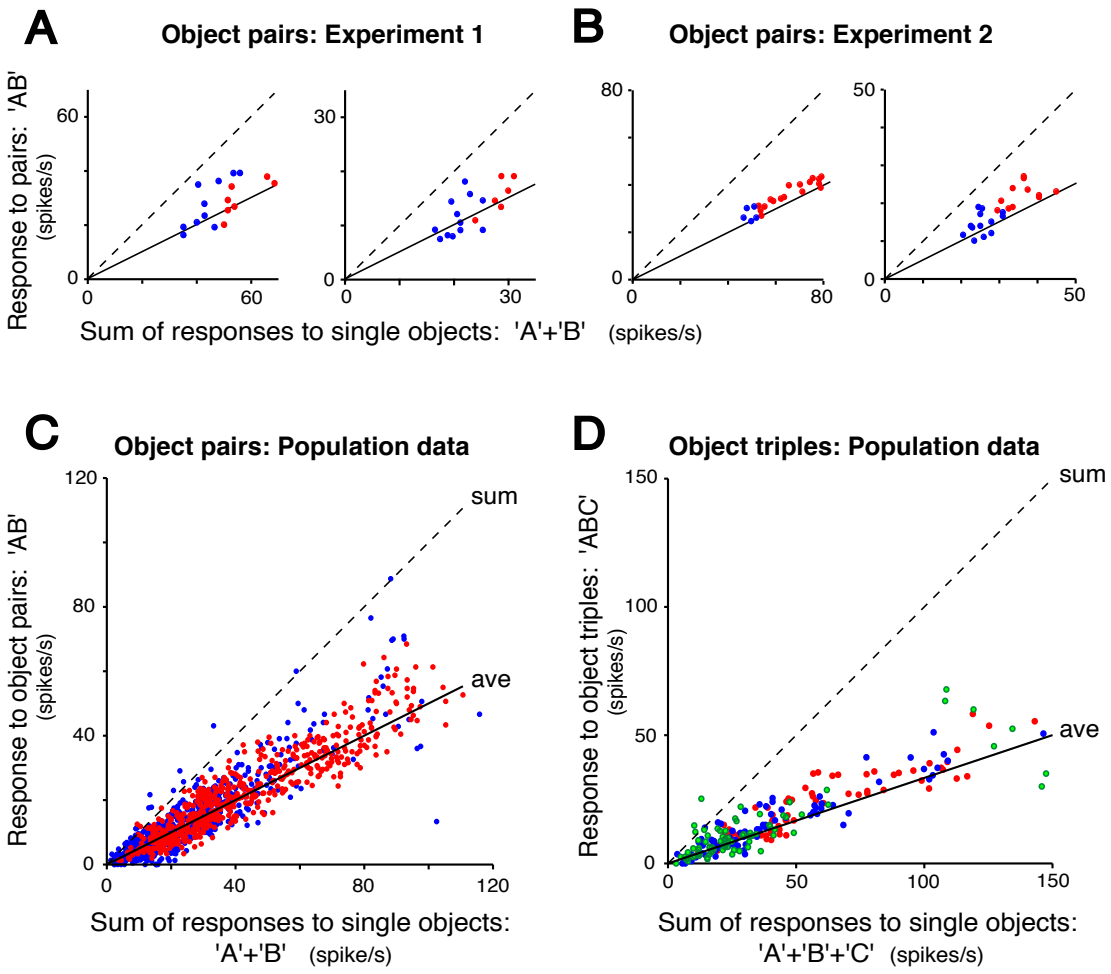


Figure 3. Responses to multiple objects as function of the sum of responses to single objects. In each scatter plot, responses to object pairs (A-C) or object triplets (D) are plotted against the sum of the responses to the constituent objects presented alone. The dashed and solid straight lines indicate, respectively, the sum and the average of the responses to single objects. The slope of the solid line is 1/2 in A-C and 1/3 in D. A. Examples data from two individual neurons recorded in Experiment1. Data in the left panel are from the same neuron shown in Figs. 2A-B. Red and blue dots refer to pairs in which, respectively, both or only one of the objects in the pair produced a response significantly higher than background rate (t-test, $p < 0.05$). B. Examples of scatter plots for two individual neurons recorded in Experiment 2. Data in the left panel are from the same neuron shown in Figs. 2C-E. Color code as in A. C. Scatter plot including responses to object pairs for the whole population of 79 responsive neurons recorded in the three monkeys. Color code as in A. D. Scatter plot including responses to object triplets for the whole population of 48 responsive neurons recorded in Experiment 1. Red, blue and green dots refer to triplets in which, respectively, three, two or only one of the constituent stimuli evoked a response significantly higher than background rate. In both the individual examples and the population data, responses to multiple objects are in very good agreement with the average of the responses to the constituent objects presented alone.

to the data was 0.55. This value is very close to the 0.5 slope expected if the responses to object pairs were the average of the responses to individual objects (solid line, referred to as the "*average* model"). Like the single neuron examples (Fig. 3A-B), this relationship was independent of the effectiveness of the less optimal object of each pair (red and blue dots are as in Fig. 3A-B). Neurons recorded in Experiment 1 were also tested with triplets of simultaneously presented objects (see Fig. 1D and Fig. 2B). Figure 3D shows that responses to triplets were also highly correlated with the sum of the responses to the constituent objects of the triplets ($r = 0.91$) and the slope of the best linear fit to the data was 0.37. This value is very close to 0.33, i.e. the slope expected if responses to the object triplets were the average of the responses to individual objects (solid line). These same analyses was repeated after normalizing all recorded responses in each neuron by the response to the neuron's most effective stimulus. This removed variance in the responses to the pairs (Fig. 3C) and triplets (Fig. 3D) that is due to differences in the range of absolute firing rates over the population of neurons. Normalized responses to pairs and triplets of objects were still well correlated with the sum of normalized responses to the constituent objects ($r = 0.58$ and $r = 0.43$ respectively for pairs and triplets) and the slope of the best linear fit to the data was very close to the slope predicted by the average model (i.e., slope = 0.44 and 0.27 respectively for pairs and triplets).

Since these previous analyses suggested that a simple *average* model might explain a great deal of the IT response to multiple objects, we sought to assess how well responses to object pairs could be accounted for by the *average* model. To do that, we determined the "goodness of fit" (GOF) of the *average* model for each recorded neuron (see Methods). The GOF provides an unbiased estimate of the percentage of data variance not due to noise ("explainable") that is explained by the model. In this case, the data variance to explain for each neuron is the variance of responses across all of the tested object pair conditions, and our goal was to assess which fraction of this variance could be explained by modeling the responses to the object pairs as the average of the responses to the constituent objects. Since the primary model under scrutiny here (the *average* model) is a function of the responses to single objects, tests of GOF of this model require modulation in the neuron's response to those objects. Therefore, we focused on neurons whose response across all tested single object conditions was highly selective (ANOVA, $p < 0.001$). These neurons were 34 of the 79 responsive neurons (15/48 for Experiment 1 and 19/31 for Experiment 2). For each of these neurons, the response R_{AB} to a pair of simultaneously presented stimuli A and B was modeled as a linear function of the sum of the responses R_A and R_B to the constituent stimuli presented in isolation, i.e. $R_{AB} = p + m\ (R_A + R_B)$. Two linear models were tested: 1) the *average* model, with $p = 0$ and $m = 0.5$ fixed for each neuron (this is simply the average of the individual responses); 2) the best linear fit to the data (with intercept p and slope m being free parameters of a Least Squares fit for each neuron). For both models, the GOF and its bootstrap standard error (SE) were computed (see Methods for details). The median GOF across the 34 tested neurons was 63.3% and 67.4% for the average and best linear model respectively. That is, allowing the two free parameters only explained an additional ~4% of variance. Figure 4 shows the distribution of the GOF values obtained for the two models. Figure 4B (last panel) also shows the distribution of the slopes m obtained by the best linear fits to the data. The median of this distribution was 0.45, which is very close to the 0.5 slope expected for the average model. The distributions of GOF values was not significantly different in Experiment 1 and Experiment 2 (Kolmogorov-Smirnov tests, $p > 0.05$). Overall, these analyses showed that, for most selective

IT neurons, responses to object pairs can be very reliably predicted as the average of the responses to the constituent objects of each pair. Indeed, the median GOF values correspond to correlation coefficients of ~0.8 (similar to the data shown in Fig. 3B, right panel).

Unfortunately, a similar analysis of GOF could not be carried out for the responses to triple objects because only 6 triplet configurations were tested in Experiment 1 (see Fig. 1D), thus yielding only 6 data points per neuron. As a result, the amount of explainable variance in the response

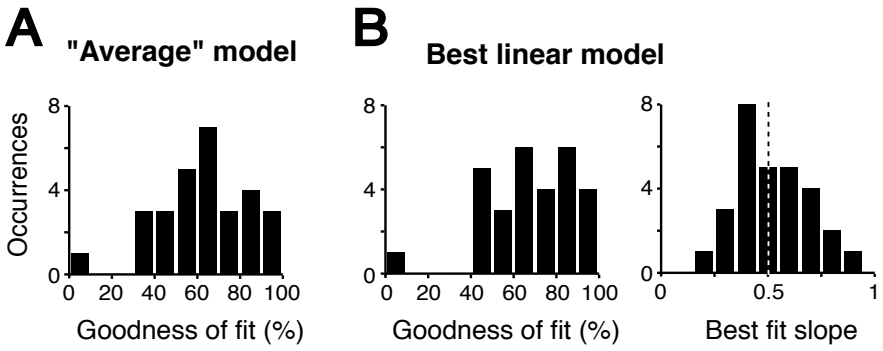


Figure 4. Goodness-of-fit (GOF) of the responses to pairs. A. GOF distribution for the average model (see Results). The GOF was computed for each of 34 highly selective neurons (see text). 29 of those 34 neurons with GOF bootstrap standard error < 40% are shown in the plot. B. GOF distribution for the best linear fit to the data (left) and distribution of the slopes resulting from that fit (right). Same neuronal population as in A. Both models explain a very large fraction of the response variance and the best linear model yields a slope distribution centered around 0.5.

to the triplets was ~1/10 of the explainable variance in the response to object pairs. That is, although our data indicate that the *average* model still holds for triple objects at the level of the IT population (Fig. 1D), the data do not have sufficient power to reliably assess the *average* model or any other model at the level of individual neurons.

Responses to single objects and pairs of objects across continuous shape dimensions

One advantage of Experiment 2 is that it allowed us to closely examine each neuron's response to pairs of objects over a continuous shape space with very similar objects (see Fig 1A, right and Fig. 2C-E). That is, we were able to find neurons that were sensitive to one of these continuous shape dimensions and measure each neuron's response along that parametric shape dimension. This, in turn, allowed us to place the neuron's preferred object in the RF and then measure the effect of adding a second object of decreasing effectiveness (when presented alone). The method

used to generate the parametric objects and screen neurons for shape selectivity is described in detail in Methods. In brief, during recordings, neurons were screened to have significantly selective responses across at least one of 42 possible shape dimensions (morph-lines) tested at fovea (ANOVA, $p < 0.05$). Objects sampled from the most selective morph-line were further tested in the top and bottom parafoveal positions (see Methods and Fig. 2D). Therefore, we were able to build tuning curves across a very selective shape dimension for each of the recorded neurons (Fig. 5) and the origin of each plot was set to be the neuron's "preferred" shape (e.g. Fig. 2C and first plot in Fig. 2D) obtained during the initial screening at fovea. Examples of such tuning curves are shown as red lines in Figure 5 (A and B) and black lines in the inset of Figure 5C. During recordings in the parafoveal locations, we also presented paired object conditions in which the preferred object obtained from the screening procedure (defined as shape distance $d = 0$) was always present, and a second object was drawn from along the tuned shape dimension (or, in some case, from another shape space, see Methods and Fig. 2E).

Figure 5A shows the data obtained from the neuron already described in Figures 2C-E and Figure 3B (left). The red line shows the neuron's response to 11 different objects sampled at increasing distances from the "preferred" object (defined as $d = 0$) along one of the shape dimensions (all presented at 1.25º above the center of gaze, see Fig. 2D). The black line shows the neuron's response to 11 object pairs conditions in which the neuron's preferred object (1.25º below fixation) was presented along with a second object (whose identity is indicated by the abscissa) at 1.25º above fixation (see Fig. 2E). The addition of the second object clearly causes the neuron's response to drop below the response of the preferred object presented alone (i.e. the black line falls below the horizontal dashed line). In fact, the response to each object pair (black line) is always in between the response to each of the constituents of the pair (i.e. in between the dashed line and the red line). At a more quantitative level, the green line shows the average of the responses to the constituent objects in each pair, i.e. the average of the dashed line and the red line. The green and black lines are almost exactly superimposed, indicating that the responses to object pairs are well predicted by a simple average model, regardless of the similarity of the paired objects.

Figure 5B shows neuronal tuning curves of another neuron (same cell analyzed in Fig. 3B, right) along a different shape dimension. Like the neuron described above, the responses to object pairs (black line) that include the "preferred" object were very close to the average of the responses to the constituent objects presented in isolation, i.e. to the average of the dashed and red curves (green line, see above). In addition, this neuron was also tested with objects sampled beyond the range of the morph-line unit distance (beyond the gray patch in Fig. 5B) and the response to pairs continued to largely track the average. Moreover, an object belonging to a different shape space (a car) was also tested both in isolation (last red point on right) and paired with the "preferred" shape (last black point on right). Even for this very dissimilar object drawn from a completely different set of shapes, the response to an object pair containing this object and the neurons "preferred" object was very close to the average of the response to each object presented in isolation (last green point on right).

Building tuning curves of the responses to single and paired object conditions (Figs. 5A-B) allowed us to test, for the neuronal population recorded in Experiment 2, if there were any consistent deviations from the average model that depended on the degree of shape similarity between

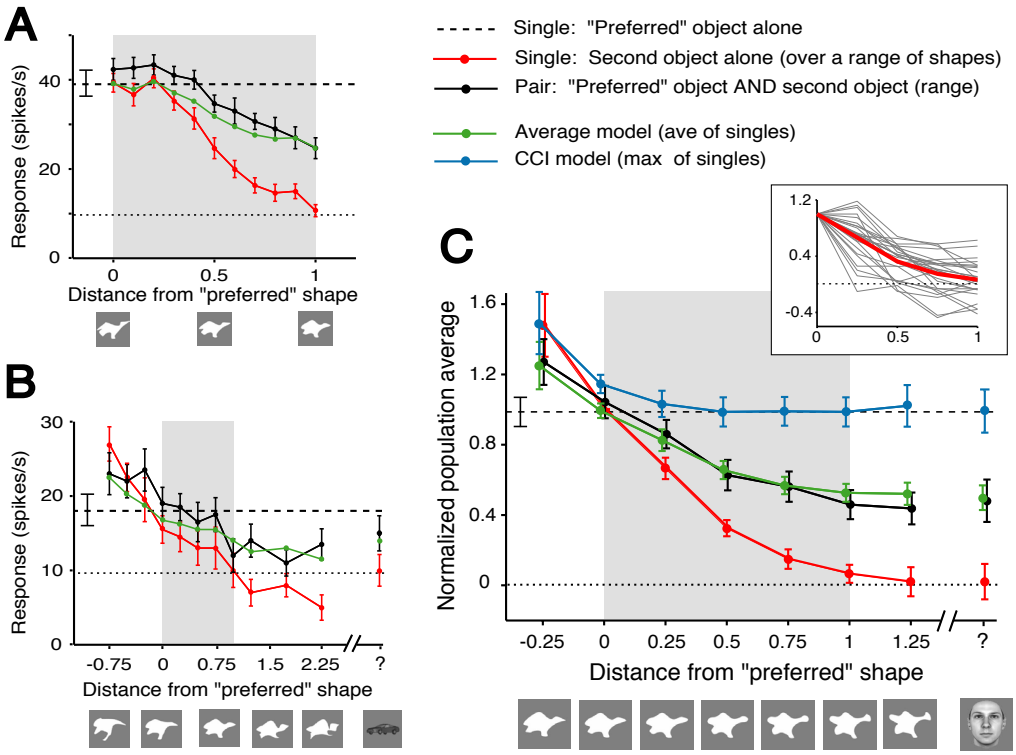


Figure 5. Tuning curves of IT neuronal responses to single objects and object pairs along continuous shape dimensions. A-B. Individual examples of tuning curves obtained for two neurons recorded in Experiment 2 (A: same neuron as in Figs. 2C-E and 3B, left; B: same neuron as in Fig. 3B, right). The abscissa is the shape distance (i.e., shape dissimilarity) within the tested morph-line (shapes corresponding to some of the tested distances are shown below each shape axis). The origin of the shape axis is the neuron's "preferred" shape obtained from the recording screening procedure (see Methods). The gray patch shows the region of shape space initially tested to obtain the preferred shape (unit shape distance, see Methods). The horizontal dotted line indicates the neuron's background rate. Morphed shapes were sampled either within the unit distance (A) or within a larger shape range (B) that included a stimulus drawn from a different shape space (data points at the far right in B). For both neurons, responses to the objects pairs (black line) are very close to the average (green line) of the responses to the constituent objects of the pairs presented in isolation (i.e., to the average of the horizontal dashed line and the red line, see Results). Error bars are SE of the mean firing rate. C. Population average of 26 tuning curves obtained from the 15 most selective neurons recorded in Experiment 2 for single and pair object conditions (see Results). These tuning curves were background subtracted, aligned to the "preferred" object (0 on the abscissa), and normalized by the response to the "preferred" object. The inset shows these 26 normalized tuning curves for responses to single objects (gray lines) and their average (red line). The red line in the main panel shows the population average of the responses to single objects and included single object conditions outside the unit shape distance (gray patch). The horizontal dashed line shows the population average of the responses to the "preferred" object of each neuron (i.e. the shape at value 0 on the abscissa). The black line shows the population average of the responses to object pairs containing both the "preferred" object and another object sampled along the abscissa. The green line shows the average model prediction (population normalized average of average model curves as in A and B). The cyan line shows the prediction of the complete clutter invariance (CCI) model (see Results). Error bars are SE of the population averages. Although different morph-lines were tested for different neurons, example shapes are shown below the abscissa from a representative morph-line. The dotted line is the background rate. Only 11/15 neurons (for a total of 18/26 responses) were tested outside the unit distance (gray patch) and contribute to the points outside this range. This plot is nearly identical when constructed from conditions where the "preferred" object was either in the best or second best RF position (top or bottom; data not shown), i.e. forcing every neuron to contribute only one tuning curve to the population average.

the objects in the pairs. To obtain a population measure of the dependence of pair responses from shape similarity, we considered the 19 highly selective neurons recorded in Experiment 2 that were included in the GOF analysis (see Figure 4) and built tuning curves for single object responses in top and bottom positions for each of these neurons, thus obtaining a total of 38 tuning curves. Because these 19 neurons were tested using different morph-lines (from different shape spaces or different axes within a shape space), the tuning curves were aligned on a single shape axis (abscissa in Fig. 5C) by choosing the origin to be each neuron's "preferred" object obtained during the screening procedure (see Methods). This "preferred" object was always one of the two objects in each object pair tested during later recordings. To get a meaningful average neuronal tuning curve, the 38 single tuning curves were screened to be both selective and largely monotonic in the unit shape distance range, i.e. within the gray patch of Figure 5C (see Methods). This resulted in a subset of 26 tuning curves recorded in 15 neurons (11 neurons contributed two tuning curves, 4 neurons contributed one tuning curve). These tuning curves were then averaged after subtracting background firing rates and normalizing by the response to the "preferred" object ($d = 0$). These 26 normalized tuning curves are shown individually in the inset of Figure 5C and the resulting population average tuning curve is shown as the red line in Figure 5C and inset. By construction, the population average (red line) falls along the abscissa as the distance from the "preferred" object is increased ($d > 0$). Note that the response typically falls to near background firing rates (ordinate = 0; dotted line) for "distant" objects sampled both within the same shape space (e.g. $d \geq 1$) and from other shape spaces (last red point on right). Note also that, although the "preferred" object ($d = 0$) was defined during initial screening, later tests sometimes included objects sampled to the "left" of the "preferred" object ($d<0$) and these tests often revealed that the response to single objects continues to increase even beyond what was taken to be the "preferred" object (i.e. red line continues to rise on the left side of 5C).

The black line in Figure 5C shows the population average of the normalized tuning curves obtained for pairs of simultaneously presented objects, in which the identity of one object of the pair was fixed at $d = 0$, while the identity of the second object spanned the tested range of shapes. Like the individual examples (Fig. 5A and B), the population average response to object pairs (black line) was intermediate between the average response to the fixed object of the pair (horizontal dashed line) and the average response to the single objects (red line). For each of the tested neurons, the response to the object pairs was modeled as the average of the responses to the constituent objects of each pair, to obtain model prediction curves as shown in Figures 5A-B (green lines). These curves were normalized and averaged to obtain the population average model prediction curve shown in Figure 5C (green line). The fact that the black line and the green line almost perfectly overlap in Figure 5C, supports two conclusions. First, the *average* model holds regardless of the similarity of the shapes composing the pairs. Second, the agreement of neuronal data to the average model prediction becomes virtually perfect when responses of even a small population of IT neurons are pooled (as done here).

Another model of responses to multiple objects

These findings clearly show that the responses of individual IT neurons are not unaffected by the addition of a second non-preferred object (i.e. they are not clutter-invariant), even when that second object produces no response on its own (see right side of plots in Figure 5A-C). Instead, the response to an effective object is predictably reduced by the presence of a less effective "clutter"

object and largely follows an average model. However, given the relevance of this conclusion for theories of neuronal representation of multiple objects (Rousselet et al., 2003, 2004) and the disagreement with some work in area V4 (Gawne and Martin, 2002), we explicitly compared the predictions of the *average* model with the predictions of an alternative model: the Complete Clutter Invariance (CCI) model. The CCI model predicts that the response to a pair of simultaneously presented objects is equal to the response of the most effective object of the pair, i.e. to the *maximum* of the responses to the individual stimuli.

The conditions used in Experiment 2 are optimized to distinguish among the CCI model and the average model because the object pairs almost always include at least one condition in which both a very effective object and a non-effective object are presented together (discussed further below). Examination of the example curves in Figure 5A and 5B clearly shows that the CCI model is not correct. In particular, the addition of a second, less effective object always causes the response to decrease below that produced by the effective object presented in isolation (the black line is well below the dashed line). To examine this for the population, CCI model prediction curves were built for each neuron included in the population averages of Figure 5C. These curves were normalized and averaged to obtain the population average CCI curve shown in Figure 5C (cyan line). The CCI model was consistently much poorer than the average model (green line) in predicting the population response to the stimulus pairs (black line). This was especially true for object conditions in which a poorly effective object was part of the pair (i.e., for $d \geq 0.5$ and for the stimulus sampled from a different shape space). Nevertheless, this is only a subset of our data and we sought to fully test the predictions of the CCI model across both experiments for all of the individual IT neurons recorded in the three monkeys.

In general, testing if responses to object pairs are better predicted by the average model (or any other model that is a weighted sum of responses to individual objects) or by the CCI model is not trivial. Although these models sound very different, the predictions of the average model and of the CCI model can be nearly identical, depending on the object conditions used to test the neurons. To illustrate this, Figure 6A shows data from our whole population of 79 responsive neurons (the same data presented in Fig. 3C), but now with the prediction of the CCI model on the abscissa. The data largely fall along the diagonal and the correlation coefficient is high ($r = 0.91$), suggesting that the CCI model does a good job in predicting the responses to object pairs. However, the reason that Figure 6A looks so clean is that a large fraction of the variance in the data is due to the differences in the firing range of the individual neurons included in the population, rather than variations due to changes in object conditions for each individual neuron. In plots like these the variance is approximately equally well explained by the CCI model, the average model (see Fig. 3C), and by any other "reasonable" model that forces responses to the pairs of objects to be near the firing range of each individual neuron.

Testing the prediction of the CCI model for each individual neuron (as done for the average model in Figs. 3A-B and Fig. 4) using measures of explained variance can also produce misleading results. If, as in Experiment 1, the identity of the most effective object in each pair is not fixed, a large fraction of the variance in the responses to pairs is due to variations in the response to the most effective object in each pair. This variance can be well explained by either the CCI model or the average model. Indeed, the median GOF index for the CCI model was 59.5% for the 15

high selective neurons recorded in Experiment 1 and included in the GOF analysis of the average model of Figure 4. However, if the neuron's most effective stimulus is paired to a set of less or at most equally effective stimuli, there is no variation in the response to the preferred stimulus in each pair and, therefore no variation in the CCI predicted response to pairs. Thus, the CCI model will fail to explain any variance in the pair responses. This is why, for the 19 high selective neurons recorded in Experiment 2 (see Figs. 4 and 5), the median GOF for the CCI model was only 3.6%.

In light of these issues, we obtained a meaningful comparison between CCI and average model by first transforming the data in the following way. Given a pair of objects A and B, with responses R_A and R_B to the individual objects and response R_{AB} to the pair AB, all three responses were normalized by dividing them by the maximum of the individual responses, i.e. MAX(R_A, R_B). As a consequence of this normalization, for each pair of objects, the normalized response to one object presented alone is equal to 1 and the normalized response to the other object presented alone is between 0 and 1. Thus, the sum $R_A + R_B$ of the normalized responses to the objects pre-

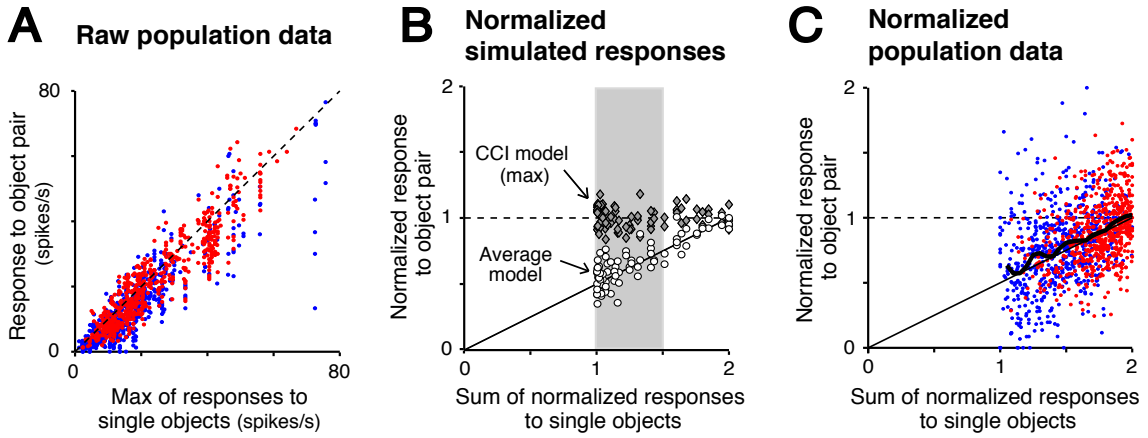


Figure 6. Comparison of the average model and the complete clutter invariance (CCI) model. A. Responses to each object pair are plotted as a function of the maximum of the responses produced by each of the constituent objects of the pair (i.e., the CCI model prediction). Data from all 79 responsive neurons are included in the plot. Color code as in Fig. 3. B. Simulated normalized responses of two model neurons, one following the "average" rule (open circles) and the other following the CCI rule (gray diamonds; see Methods). In the first case, the neuron's response to object pairs was modeled as the average of the response to the constituent objects of the pair; in the latter case as the maximum of the constituent responses. Responses to each object pairs and the two constituents of that pair were normalized by the maximum of the latter two. As a consequence, the sum of the normalized single object responses (in abscissa) ranges from 1 to 2, while the normalized responses to pairs cluster around the solid line with slope = 0.5 for the average model simulated neuron and around the dashed line with slope = 0 for the CCI model simulated neuron. The gray patch shows the range in which the predictions of the two models can be most easily discriminated. C. Normalized responses for the whole population of 79 neurons (i.e. normalized as in B). Color code as in Fig. 3. The heavy black curve is the average response to object pairs as function of the sum of individual responses. The average is computed in a running window of size 0.1 shifted in consecutive step of size 0.05.

sented alone is between 1 and 2. It is easy to show that, once data are transformed as described above, the predictions of the CCI and of the average model become distinguishable, regardless of the set of objects used to test them. This is shown in Figure 6B, where normalized responses to pairs (R_{AB}) are plotted against the sum of normalized single responses ($R_A + R_B$) for two different simulated neurons (one following a CCI rule and one following an average rule; see Methods). The scatter plots in Figure 6B shows that data points generated by the *average* model neuron (empty circles) line up along the straight line with slope 0.5. The data points generated by the CCI model neuron (gray diamonds) line up along the line with slope 0. When responses to the object pairs were modeled as a linear function of the sum of individual responses, i.e. $R_{AB} = m (R_A + R_B)$, with variable slope m, then scatter points lined up along straight lines with slope equal to m (simulation data not shown).

Data from the population of 79 responsive neurons recorded from the three monkeys were normalized as described above and then plotted in the population scatter plot shown in Figure 6C. The data points were scattered around the straight line with slope 0.5 (solid line) and a running average of the responses to pairs as function of the sum of individual responses was almost superimposed to the slope 0.5 line (heavy black line; see caption for details). This shows a virtually perfect agreement of neuronal data to the average model prediction when responses from the whole population of recorded neurons were averaged. Because the average slope in Figure 6C remains at 0.5 across the entire range of possible abscissa values (1.0 – 2.0), this shows that agreement did not depend on the effectiveness of the individual objects, thus confirming the conclusions from the previous section (see Fig. 5).

To further compare the average model and the CCI model, we focused on the stimulus conditions under which the predictions of the *average* and CCI model are most disparate. Specifically, as suggested by a previous study (Gawne and Martin, 2002), we considered only object pair conditions in which the less effective objects in isolation evoked a response that was less than half the response evoked by the more effective object. In the transformed data described above, this corresponds to data that have abscissa values <1.5. We computed the median response to pairs (R_{MED}) for each neuron across this subset of conditions ($1<[R_A+R_B]<1.5$; gray patch in Fig. 6B). If the pair responses follow the *average* model and the data were uniformly distributed across the 1-1.5 interval, then the R_{MED}s for the recorded neurons should be distributed around 0.625 (in fact, the data were not evenly distributed over this interval so the *average* model predicted a R_{MED} distribution centered around 0.68). On the other end, if the pair responses follow the CCI model, the R_{MED}s should be distributed around 1.0. The observed median of the R_{MED}s across a population of 64/79 neurons recorded in the three monkeys was 0.7 (mean=0.7), i.e. very close to the value predicted by the *average* model (15 neurons were excluded from the analysis because they had no points in the interval $1<[R_A+R_B]<1.5$). Put another way, this shows that, on average, the response of an IT neuron to an effective object is reduced by 30% when that effective object is presented with a "less than half" effective second object. At an individual neuron level, 43 of the 64 neurons had median responses to these object pairs that were reduced by at least 20% (relative to the response to the preferred object presented alone). We also observed that ~12% of the neurons (8/64) had responses to these object pairs that were reduced by less than 5% and might thus be taken to be consistent with the CCI model. Overall, however, the vast majority of IT neurons give responses to pairs of objects that are far from the CCI model prediction (see Discussion).

Response to objects pairs as a function of RF sensitivity

The present study was not designed to explicitly test the dependence of responses to objects as a function of their RF position in that the spatial separation of objects in the RF was not systematically varied and only two or three RF locations close to the center of gaze were tested. However, since IT neuronal RFs are not all centered at the same retinal position, have a broad range of sizes (Op De Beeck and Vogels, 2000), and can often be small relative to the separation of our objects (Op De Beeck and Vogels, 2000; DiCarlo and Maunsell, 2003), we used these RF variations to ask if there was any relationship between the averaging behavior described above and position in the RF. In particular, we might not expect the response to a pair of objects to be the average of the responses to the constituent objects if one of those objects was presented very far outside the RF (Missal et al., 1999), but we wondered if we might detect some breakdown in averaging behavior when one of the objects was near the edge of the RF.

To examine this, we first defined the sensitivity of the RF at each tested position as the average response to objects that were effective in at least one position (i.e., eliciting a response significantly higher than background; t-test, $p < 0.05$). We then examined the *average* model as a function of the relative effectiveness of the two RF positions. Specifically, for each tested object pair condition, we computed the response to the object pair as a fraction of the sum of the responses to the constituent objects, i.e. the ratio $R_{AB} / (R_A + R_B)$. As described above, this value tends toward 0.5 (*average* model) when all our data are considered together. Figure 7 shows this ratio as a function of the relative RF effectiveness of the two positions (the red line is a running average). Two points can be taken from Figure 7. First, as expected based on the placement of our objects and the distribution of IT RF sizes (Op De Beeck and Vogels, 2000), for most neurons both objects were well within the RF (relative effectiveness values are all > 0%), but, for some neurons, one of the tested positions was near the edge of the RF (i.e. 20% RF effectiveness). Second, over this range of RF sensitivity conditions, we see only a very slight trend away from averaging (and towards no effect of the

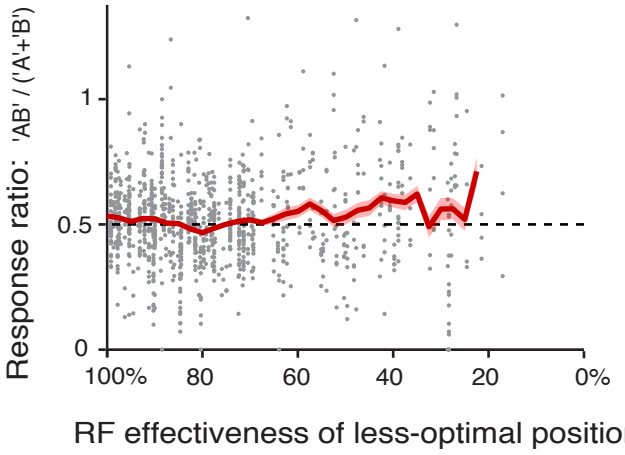


Figure 7. Agreement between responses to object pairs and prediction of the average model as function of the receptive field (RF) sensitivity. The abscissa shows the RF effectiveness of the less effective position occupied by one of the two objects. The ordinate is the ratio of the responses to object pairs to the sum of responses to the constituent objects. Each gray point is one pair condition from one neuron and all 79 responsive neurons are included. The solid red curve line is the average in a running window of size 10% shifted in consecutive steps of size 2.5%. The red shaded region is ±1 SE of the running average. The horizontal dashed line shows the ratio predicted by the average model (0.5).

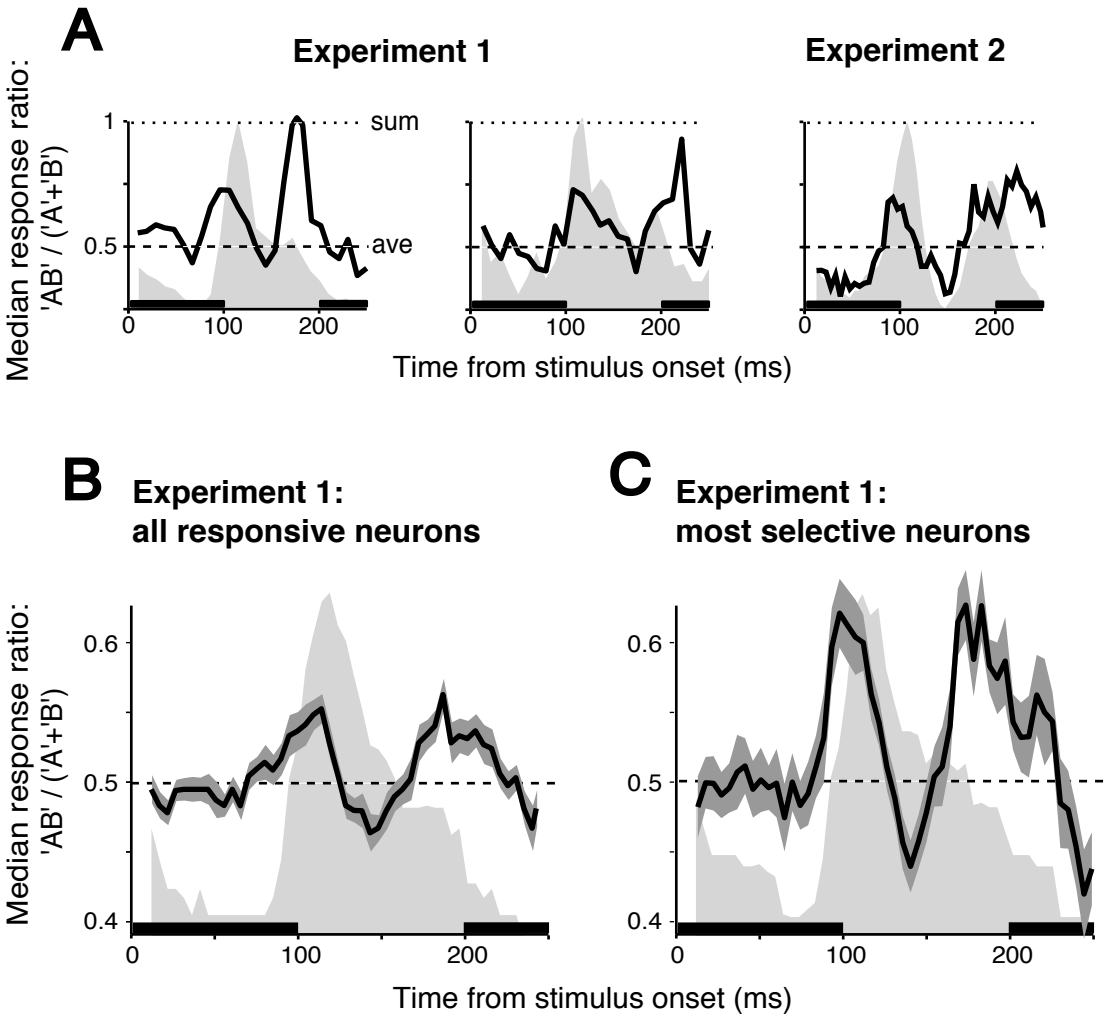


Figure 8. Dynamics of the response normalization. A. Time course (solid line) of the median ratio between responses to object pairs and sum of the responses to the constituent objects of the pairs, for two individual neurons recorded in Experiment 1 (left and middle panels) and one neuron recorded in Experiment 3 (last panel). Neuronal responses are computed in overlapping time windows of 25 ms shifted in time steps of 5 ms. The light gray background shows the time course of the median response to object pairs for each neuron. The heavy bars along the abscissa show timing and duration of stimulus presentation (see Fig. 1E) and all calculations are based on single objects or object pairs presented at time zero. The dashed line is the prediction of the average model without dynamics. The dotted line is the prediction of a sum model (i.e. a model in which the response to a pair of objects is the sum of the responses to the constituent objects). B. The solid line is the median of the ratios between responses to object pairs and the sum of responses to the individual objects, median over all object pairs tested across the whole population of 48 responsive neurons of Experiment 1. The shaded regions are ±1 SE of this median (the SE was computed by bootstrap re-sampling of the ratios). The light gray background is the median over the responses to object pairs across the 48 neurons. C. Same as in B, but obtained for the subpopulation of 10 highly selective neurons recorded in Experiment 1 (i.e., neurons whose responses to all single stimulus conditions were significantly different at p = 0.0001, ANOVA).

second object) as we approach the edge of the RF. This trend is consistent with the reduction of response suppression produced by the less effective shape in a pair as a function of its distance from the more effective shape (Missal et al., 1999).

Dynamics of the response to object pairs

In the previous analyses, IT neuronal responses were computed in a fixed time window of 100 ms (i.e., between 100 ms and 200 ms from the stimulus onset; see Fig. 2). Because our observations appear consistent with some sort of normalization mechanism, we wondered if we could detect any temporal structure in the pattern of responses that might be consistent with such a mechanism. To do this, we first computed the average firing rate of the recorded neurons in small time bins (25 ms width) shifted in consecutive time steps of 5 ms. Then, for each neuron and each time bin, we computed the median ratio between responses to object pairs and the sum of responses to the individual objects, i.e. $R_{AB} / (R_A + R_B)$, median over all object pairs tested for each neuron. As described above, this ratio tends toward 0.5 (*average* model) when data are considered over our standard 100-200 ms post-stimulus interval. The resulting time course of the median ratio is shown for two neurons recorded in Experiment 1 and one neuron recorded in Experiment 2 (Fig. 8A). For comparison, the time course of each neuron's median response to the object pairs is also shown in each panel (light gray background). Figure 8 shows that, for all three neurons, before the onset of the response (i.e., up to ~100 ms after stimulus onset) the median ratio fluctuates around 0.5. This is expected because the background rate during presentation of single objects and pairs of objects should be approximately the same. Then, at the beginning of the neuronal response (~100 ms post-stimulus onset), the median ratio increases *above* 0.5. The peak ratio then decreases, reaches a minimum (below 0.5) around 150 ms from the stimulus onset and then reaches a new peak (above 0.5) around 200 ms from the stimulus onset.

This temporal pattern suggests that, at the onset of the neuronal response, responses to object pairs are *above* the average of the individual responses to the constituent stimuli, i.e., in the direction predicted by the sum. This pattern was found for many neurons recorded in Experiment 1 and for some neurons recorded in Experiment 2. To examine this across the recorded neuronal population, we computed the time course of the median ratio between responses to object pairs and the sum of responses to the individual objects, median over all object pairs tested across the whole population of 48 responsive neurons of Experiment 1. The resulting curve (Fig. 8B, solid line) showed dynamics very close to that observed in the individual neurons except that the peaks and trough were smaller. However, when only the 10 most selective neurons of Experiment 1 were taken into account (see Fig. 8 caption), the peaks and the trough were much more pronounced (Fig. 8C, solid line). Overall, these observations suggest that a short time lag is involved in the mechanisms underlying the *average* effect and thus hint at the possibility that competitive normalization may be the mechanism underlying the *average* response to multiple objects described above (see Discussion).

The temporal pattern shown in Figure 8 was less pronounced in Experiment 2, although it was observed (e.g. Fig. 8A, last panel). When the 31 responsive neurons of Experiment 2 were considered (as in Figure 8B), the resulting curve had a time structure similar to that shown in Figure 8B (not shown). However, the first peak at the time of response onset (~100 ms) was much less prominent. The absence of a clear peak may have been due to more frequent saturation of neuro-

nal responses in Experiment 2 because, unlike Experiment 1, it involved an effective object in all paired conditions (see Fig. 2E).

Possible role of attentional shifts

As a final step, we sought to understand if the *average* effect described throughout this study could have resulted from effects of visual attention. When two objects are present and the monkey is cued to attend a specific visual field location (Connor et al., 1997; Reynolds et al., 1999) or a specific target object (Treue and Maunsell, 1996; Chelazzi et al., 1998), neuronal responses in the ventral visual stream (including IT) move toward the response elicited by the attended object, as if that object were presented alone. Thus, if one of the two objects in each pair were attended on each presentation of the pair, and the choice of the attended object were random across the 10-30 trials in which the pair is tested, the mean response over all trials could look very much like the average of the responses to the constituent objects presented alone. Although our presentation conditions (100 ms stimulus duration) are likely far too rapid for attentional shifts *during* a single presentation of a pair, if the animal's attention were directed toward one position for approximately half of the trials and the other position for the rest, attentional shifts might explain the *average* effect.

This hypothesis makes the explicit prediction that the distribution of responses across the 10-30 presentations of each object pair contains responses that are drawn more or less equally from the distributions of responses to the two constituent objects. This, in turn, predicts that the distribution of responses to the pair should be very broad (and possibly bimodal), especially for cases in which one object is very effective and the other is non-effective. The broadness of spike discharges is typically quantified by the Fano factor, i.e. the ratio of the variance of the average spike count and its mean. The distribution of spike counts elicited by repeated presentations of a single visual stimulus is well known to approximately follow a Poisson distribution with mean equal to the average spike count and thus have a Fano factor of ~1 (Softky and Koch, 1993; Shadlen and Newsome, 1994; Rieke et al., 1997; Shadlen and Newsome, 1998). If attention were not a factor, the distribution of responses to pairs of objects should be the same as that produced by single objects and should have a Fano factor of ~1. Indeed, for the 79 responsive neurons, the average Fano factor obtained for pair conditions (1.05) was not significantly higher than the average Fano factor obtained for single conditions (1.13; one-tailed unpaired t-test, $p = 0.98$). Furthermore, we simulated response distributions to object pairs that should have resulted from the random allocation of attention by randomly sampling 10-30 responses from the distributions observed for each of the two constituent objects. The average Fano factor of these simulated pair responses was 1.24, i.e. ~ 20% higher than the Fano factor obtained from the actual pair responses. This difference was highly significant (one-tailed paired t-test; $p < 0.001$). Overall, these comparisons strongly suggest that the observed *average* effect described throughout this study cannot simply by explained by alternating attention shifts.

Discussion

In pursuing an understanding of the mechanisms underlying visual recognition in cluttered, real-world scenes, the goal of the present study was to systematically examine IT neuronal responses in limited clutter conditions, i.e. with multiple objects present, using two complementary ex-

perimental paradigms. The approach of Experiment 1 was to test the exact same visual object conditions across an unbiased sample of IT neurons. This approach produced an IT population that was unbiased with respect to the objects tested for each neuron, but did not maximize the response range for each neuron. The complimentary approach of Experiment 2 was to optimize the tested object conditions for each neuron to produce maximal selectivity across a continuous shape dimension and, as a consequence, may have increased the probability that each neuron was involved in the representation of the objects tested.

Our results show that, across this wide range of stimulus conditions a large fraction of the explainable variance (~63%) in the responses to object pairs was accounted for by the average of the responses to the constituent objects (i.e., *average* model; Fig. 4A). Because of the consistency across the population of IT neurons, the *average* model becomes virtually perfect when responses of even a small population of neurons are pooled (Fig. 5C). One corollary of the *average* model is that the IT response to a pair of objects depends largely on the relative effectiveness of each of the constituent objects in driving the neuron, and that it does not much matter if that effectiveness is altered by changing the object identity (Fig. 5) or the RF position at which that object is presented (Fig. 7). Another corollary is that objects that are completely ineffective in driving a neuron when presented alone powerfully reduce the neuron's response when paired to very effective objects (at least for the conditions tested here, see below). As such, it is clear that most IT neurons do not have complete clutter invariance (CCI), i.e. do not have responses that are completely independent of the presence of a less effective object (see Fig. 5C). This was confirmed by directly comparing the average model and the CCI model (Fig. 6).

Previous studies of multiple stimuli interaction in the ventral visual stream

Consistent with previous investigators (Sato, 1989; Miller et al., 1993; Rolls and Tovee, 1995; Missal et al., 1997; Chelazzi et al., 1998; Missal et al., 1999; Sheinberg and Logothetis, 2001), our study found that IT responses to very effective stimuli are typically reduced by the presence of less effective "clutter" stimuli. In particular, we found that, on average, responses of IT neurons to an effective stimulus were decreased to ~ 70% when a "less than half" effective stimulus was also presented, which is very close to the magnitude of suppression reported by Rolls and Tovee (1995) and Missal et al. (1999) for similar object pairings. Although not systematically described, there are hints of an *average* rule in previous IT studies. For example, Miller et al. (1993) found a correlation between the degree of response suppression produced and the relative effectiveness of the RF location at which a second stimulus was presented – implying that the degree of suppression depends on the neuron's response to the second stimulus presented alone.

On the other hand, in apparent disagreement with the results found here, Missal et al. (1999) did not find correlation between responses to object pairs and the sum of the responses to the constituent objects. This may have resulted from failure to probe IT neurons over a sufficiently large range of stimulus effectiveness. In particular, because a very effective shape was always paired with a poorly or non-effective shape, there would have been little or no variance in the sum of the responses to individual objects (abscissa in Fig. 6C), making it difficult to reliably detect the *average* effect. Thus, we believe that the data of Missal et al. do not contradict the *average* effect, but that the systematic relationship between responses to multiple and single stimuli reported in our study may not be revealed under more limited testing conditions.

Previous studies also appear to disagree on whether and how response suppression depends on the identity of the second, less-effective object. Miller et al. suggested that the amount of suppression did not depend on that identity, while Missal et al. found the opposite for 50% of neurons. Our results indicate that the answer depends not on the object identity per se, but mainly on the amount of activation produced by that object when present alone (with important caveats for completely non-effective objects, see below). That is, on average, objects that produce the same response when presented alone produce the same amount of response suppression when paired to an effective object (compare, for instance, the points at the far right of Fig. 5C). This does not rule out the possibility of shape-dependent deviations from the *average* model at the level of single neurons (see example in Fig. 5B). However, our results suggest that any such deviations would be averaged out by pooling responses of an even small population of IT neurons.

Our results are in good agreement with previous investigations of the responses to multiple stimuli in areas V2 and V4 (Reynolds et al., 1999). Reynolds et al. found that responses of V2 and V4 neurons to pairs of simultaneously presented stimuli can be reliably modeled as a weighted sum of the responses produced by the two stimuli in isolation. They also found that, when the attention of the monkey was not directed to any stimulus in the pair, V2 and V4 neuronal population responses to stimulus pairs tended to follow an *average* model. Thus, we speculate that the same interaction mechanisms engaged by multiple visual stimuli may operate across different stages of the ventral visual stream to produce the observed averaging effect.

This conclusion does not fit with the results of a recent study (Gawne and Martin, 2002), in which V4 neuronal responses to stimulus pairs were similar to the maximal response of the constituent stimuli presented alone (CCI model). To explore this possibility in IT, we directly compared the *average* model and the CCI model (Fig. 6), and we also specifically examined object pair conditions with constituent responses in the same range tested by Gawne and Martin. In both analyses we found a clear agreement of population data to the *average* model (Fig. 6C), and that only a small percentage of neurons (~10%) have responses consistent with the CCI model. Nevertheless, it is important to realize that the *average* model cannot hold for all non-effective objects, e.g. if distractor objects were presented far outside the RF or with attributes that do not penetrate the visual system. That is, depending on the non-effective distractor objects used, the response to object pairs could appear to follow the *average* model, the CCI model, or something in between. Thus the apparent inconsistency of the results of Gawne and Martin with previous investigations in V4 and with our study in IT may be due to a higher degree of stimulus separation (as pointed out by the authors) or the possibility that some non-effective stimuli may have been of too high spatial frequency at the presented eccentricity to penetrate the visual system.

Possible neural mechanisms underlying the average model

The *average* model presented in this paper is only a descriptive "model" and leaves open the question of underlying mechanisms. Reynolds et al. (1999) proposed a mechanistic implementation of the "biased-competition model" (Desimone and Duncan, 1995) that can explain the weighted average of responses to constituent stimuli of a pair in V2 and V4. That model assumes that each object in a pair activates a separate population of afferents to the neuron and a normalization factor proportional to the total synaptic input rescales the neuron's response to the pair. A similar mechanism by which the output of each IT neuron is normalized by its total synaptic

drive could also explain the *average* effect we observed. The *average* effect could also arise if the output of each IT neuron was normalized by the total spiking activity of a broad population of IT cells – similar to a class of divisive normalization (or gain control) models proposed to explain nonlinear behavior of neurons in early visual stages (Heeger, 1992; Heeger et al., 1996; Carandini et al., 1997; Schwartz and Simoncelli, 2001; Cavanaugh et al., 2002) and response re-scaling in area MT (Recanzone et al., 1997; Britten and Heuer, 1999; Heuer and Britten, 2002). It is also possible that feedforward mechanisms leading up to IT could produce the average rule. For example, a biologically constrained computational model that can produce key selectivity and invariance properties of IT neurons may also provide an explanation of the *average* effect in IT, even though it was not explicitly constructed for that purpose (Riesenhuber and Poggio, 1999a, b; Poggio and Bizzi, 2004). Our finding that the initial response to object pairs is often higher than the average of individual responses (Fig. 8) is suggestive of an initial linear-like sum of synaptic drive, which is rescaled by a gain mechanism after a short delay.

Relevance of the average rule for object representation

Whatever mechanism underlies the average rule, it will be crucial to understand its implications for object representation in IT. First, it should be noted that the *average* effect does not change the preferred objects of IT neurons but rescales their tuning properties (see Fig. 5), consistent with the preservation of selectivity profiles of IT neurons found in studies using natural visual scenes (Sheinberg and Logothetis, 2001; Rolls et al., 2003). Nevertheless, the presence of a second object clearly changes each neuron's magnitude of response to its preferred object and thus, at first glance, suggests that the *average* effect will negatively impact recognition of the preferred object. However, while no impact of a second object on the response to a preferred object may seem to be a desirable property of individual IT neurons for robust object recognition (Rousselet et al., 2003, 2004), it is not obvious that such a property is necessary or useful when populations of IT neurons are considered. Indeed, IT neurons following the *average* rule carry information about the identity of both objects in a pair that is lost by the CCI rule. Therefore, a population of IT neurons following the *average* rule might allow the simultaneous representation of multiple visual objects.

Chapter 4:
Can Inferotemporal Cortex Simultaneously Represent Multiple Objects?

The primate inferotemporal (IT) cortex is widely believed to code for object identity in a manner that is largely invariant to object position. However, primates function in complex visual environments and must be able to recognize both the identity and position of objects, even when multiple objects are present. While position invariance would seem desirable for representing isolated objects, it poses computational problems when viewing multiple objects, because it becomes unclear how to combine information about "what" object is present with information about "where" it is. This problem is compounded by the fact that IT neurons show nonlinear, suppressive responses when multiple objects are present – suggesting that representation of one object may interfere with representation of another object when both are present. However, by studying IT responses to single and multiple objects, we found that populations of IT neurons contain significant information about both the identity and position of each object *in parallel*, and that this information requires only simple mechanisms to read out. This shows that reasonably sized IT populations contain enough information to avoid representational interference when multiple objects are present and suggests that imperfect invariance in single-unit responses may not be shortcomings, but desirable properties for real-world recognition.

Introduction

Visual information flows through two relatively distinct processing streams in cortex: a ventral "what" stream, thought to be more involved in determining object identity, and a dorsal "where" stream, thought to be more concerned with spatial aspects of vision such as motion and object position (Ungerleider and Mishkin, 1982). In its most extreme conception, this segregation leads to one form of the infamous "binding problem," in which information about object identity and object position must somehow be "bound" back together to correctly interpret a scene. While many solutions to this problem have been proposed (e.g. Treisman, 1999; von der Malsburg, 1999), relatively few authors have considered the possibility that the ventral stream might, by itself, be capable of simultaneously representing both object identity and position (Edelman and Intrator, 2003; Rousselet et al., 2004). Here, we ask empirically if the culmination of the

primate ventral stream, the inferotemporal cortex (IT), is capable of simultaneously representing both "what" and "where" when multiple objects are present.

The suitability of a neural representation for jointly representing both object identity and position of multiple objects depends on a number of factors. The simplest and most obvious requirement for such a joint representation is the presence of position information in the first place. Put another way, position invariance cannot be absolute. Existing electrophysiological data with single objects are consistent with this idea: although the position "invariance" properties of IT neurons are often highlighted (Gross et al., 1969), IT neuronal responses exhibit only a relative form of position invariance. While each IT neuron responds best to its preferred objects at its preferred position in the visual field (its receptive field center, RF), the overall magnitude of response decreases as the preferred object is positioned at increasing distances from the RF center (Op de Beeck and Vogels, 2000; DiCarlo and Maunsell, 2003). In other words, even though IT RFs can be quite large, they do not span the entire visual field and are not uniformly responsive across their extent, and thus they may convey substantial position information in addition to object identity information.

However, even if individual IT neurons carry some position information about objects viewed in isolation (as in the above studies), this provides little insight into how IT represents *multiple* objects, or whether it can represent those objects in *parallel*. Answering these questions requires an understanding of how IT neurons respond when multiple objects are present in their RFs, and the reality of IT responses to multiple objects is far from straightforward. In essentially all studies of IT responses to multiple objects, responses to preferred objects are significantly altered by the presence of non-preferred objects (Miller et al., 1993; Missal et al., 1999; Rolls and Tovee, 1995; Zoccolan et al., 2005) or complex backgrounds (Rolls et al., 2003; Sheinberg and Logothetis, 2001). In the face of such non-linear interference effects, it is difficult to see clearly how IT could accurately represent multiple objects simultaneously.

Nevertheless, it is extremely difficult to intuit the behavior of neuronal populations from the response properties of individual neurons. Indeed, when considered at the population level, such nonlinear effects may not constitute "interference," but may instead reflect useful coding schemes (Wainwright et al., 2002, see Discussion). In either case, what IT cortex does and does not represent at the population level remains an important, open, empirical question. The goal of the present study was to ask if the IT population carries a simultaneous representation of *multiple* objects *and* their positions.

Results

We recorded the responses of a population of monkey IT neurons (n=68) to a common set of visual stimuli, including single objects and combinations of those objects (pairs and triplets, Figure 1). We then used linear discriminant classifiers (Fisher, 1936) as a simple, unbiased means to ask what information is *directly* conveyed by the IT population. Note that we are not simply asking if *any* object information is conveyed by the population (e.g. we are not simply assessing Shannon Information; Shannon, 1963; Tovee et al., 1993). Instead, because each linear discriminant

does nothing more complex than perform a weighted sum (with a threshold; Gochin, 1994; McCulloch and Pitts, 1943), this method allowed us to specifically assess information in the IT population that could be *directly* extracted by simple mechanisms that roughly parallel mechanisms available to real neurons that receive inputs from IT (see Experimental Procedures). In other words, the amount of information that can be extracted using this method is a measure of information that is carried in the IT population in a format readily available to downstream neurons.

Since IT is widely thought to be involved in the recognition of object identity irrespective of retinal position (e.g. Schwartz et al., 1983; Logothetis et al., 1995; Ito et al., 1995; Op de Beeck and Vogels, 2000), we first built linear discriminants to determine how well the IT population can report object identity, regardless of position. We began by examining the simple situation in which each image contained just one object (three possible objects, three possible positions within 2º of the center of gaze; see Experimental Procedures for details). We found that the IT population could support this task well above chance (mean: 69.1%; $p << 10^{-6}$; chance = 50%; for a discussion of chance levels and p-values, see the Experimental Procedures section). This performance (and all the performance results listed in this paper) is that which could be achieved using the IT population spike response data on a *single trial* (100 ms image presentation, 100 ms of response data), and those data were never previously seen by the linear discriminant classifier. Beyond providing quantification, this result is perhaps not surprising given existing data showing position tolerance of IT neuronal selectivity with isolated objects (Schwartz et al., 1983; Logothetis et al., 1995; Ito et al., 1995; Op de Beeck and Vogels, 2000).

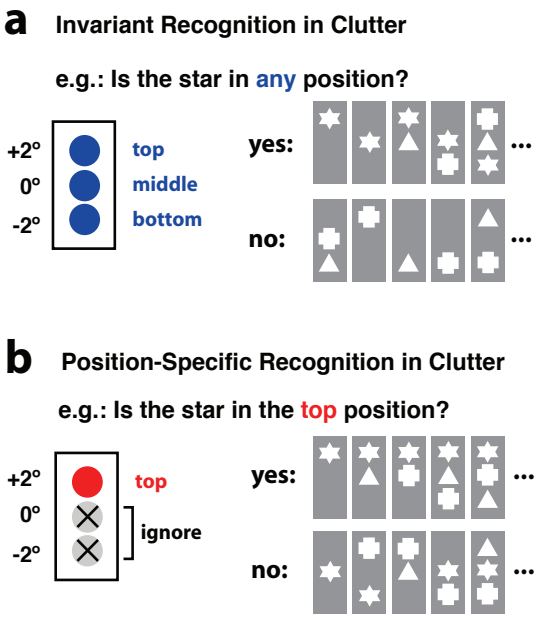


Figure 1. a) Position-invariant recognition in clutter. Three objects (star, triangle, and cross) were shown in isolation at each of three positions in the visual field (center of gaze, 2o above and below the center of gaze). The same three objects were also presented in all pair-wise and triplet-wise combinations using the same three positions (see Supplementary Methods). We built linear dis-criminant classifiers to ask the IT population "questions" about the identity of the presented object(s) An example of one such question is shown, along with some of the visual conditions that served as positive and negative examples ("yes" or "no"). We asked this particular linear dis-criminant to determine if the star object was present in any position, while ignoring the presence of other objects. b) Position-specific recognition in clutter. One example position-specific ques-tion is shown, along with several positive and negative examples. In this case, the classifier was asked to determine if a star was present at the top position, irrespective of the presence of objects at other positions.

Next, we considered a more complex situation in which we also included images containing multiple objects. Specifically, the image set included images of single objects (above), as well as images of two or three objects (see Fig. 1). Again, we built linear discriminants to ask if the IT population can report object identity, regardless of object position. This is exactly the same as

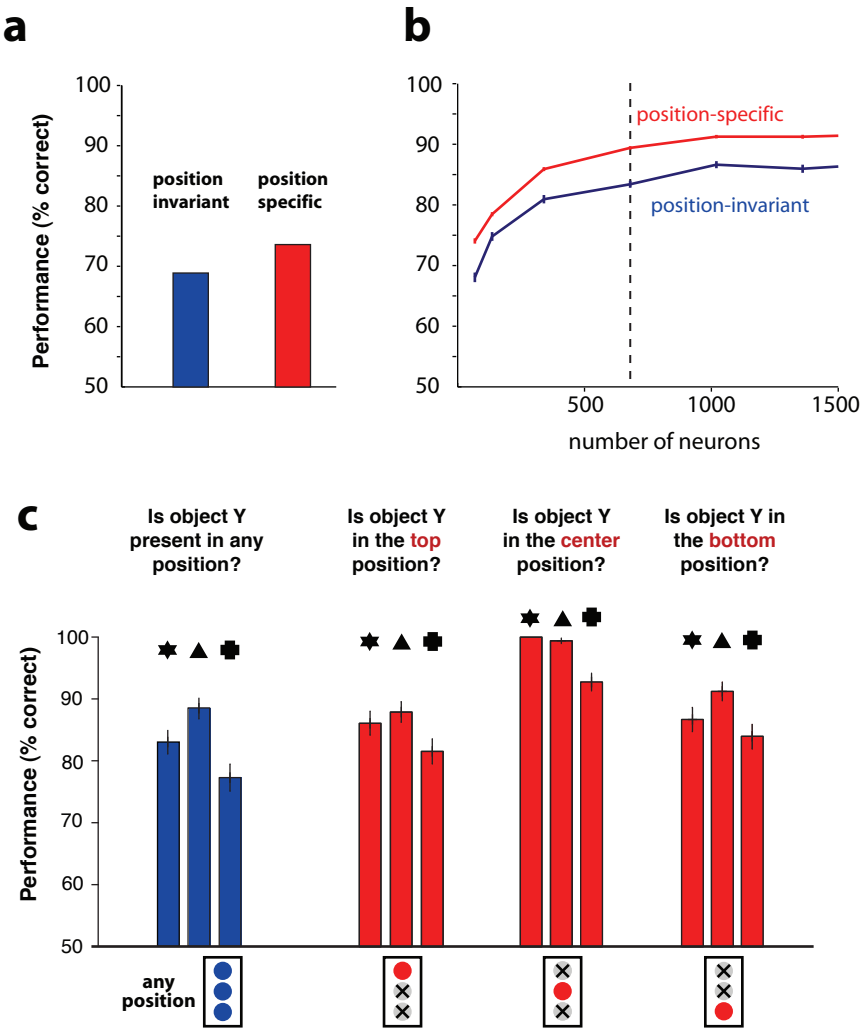


Figure 2. a) Object recognition performance in clutter. Average cross-validated performance is shown for the recorded population of neurons (n=68) for both position-invariant classification questions (blue bar; see Figure 1a) and position-specific questions (red bar; see Figure 1b). In both cases, performance is highly significantly above chance (50%, see Experimental Procedures). b) Av-erage performance as a function of synthesized population size (see Experimental Procedures) for position-invariant classification questions (blue line) and position-specific questions (red line). c) A detailed breakdown of performance for a synthesized population of 680 neurons (population size corresponds to the dotted line in Figure 2b). Blue bars represent position-invariant questions (of the sort posed in Figure 1a); red bars represent position-specific questions (of the sort posed in Figure 1b).

the previous test except that it now assesses performance in the face of other "distractor" objects (i.e. multiple object conditions, see Fig. 1). Because we have recently shown that the presence of such multiple objects strongly suppresses the responses of these *individual* IT neurons (Zoccolan et al., 2005; see Methods), one might predict poor IT population performance in this test. However, we found performance well above chance (mean: 68.9%; $p << 10^{-6}$; Fig. 2a), and only slightly degraded from that observed with single objects (c.f. 69.1% above). This shows that the IT population contains information to support position-invariant object identification, even in the face of visual clutter.

Finally, to directly address the question of representing multiple objects simultaneously, we asked if the IT population could report object identity at multiple positions in parallel. To do this, we used the same set of images (single and multiple object conditions) and response data, and we built linear discriminants to perform the same object identification task *at each individual position* (see Figure 2). At each of the three positions tested, these classifiers performed as well as, or better than, the position-invariant classifiers (mean: 73.6%, above chance at $p << 10^{-6}$; Fig. 2a), indicating that it is possible to determine stimulus identities *at particular positions* from IT population responses at least as well as it is to extract identity per se. This means that downstream neurons could, in parallel, reliably report the identity and position of each object in the image (at least up to the limited clutter conditions tested here, see Discussion).

One advantage of the quantitative approach we have take here is that is allows us to determine the amount of directly available information for such tasks under a number of different assumptions about how the representation is "read-out" by downstream neurons. In particular, it is well known that population size can strongly influence the reliability of signals and thus increase the total amount of conveyed information. It is also known that cortical neurons can receive a number of synaptic inputs (~10,000; Braitenberg, 1978) that is much larger than the number of IT neurons that can reasonably be recorded with current techniques. Thus, we used the linear discriminant approach to characterize how the amount of directly available information would scale with increasing numbers of IT neurons. To do this, we synthesized larger populations of Poisson-spiking neurons from the response profiles of the measured IT population. This procedure does not assume any stimulus selectivity that was not already in the population (because all synthesized neurons are copies of one of the original 68 neurons), but it does allow for moderate amounts of pooling to overcome the high trial-to-trial variability of cortical neurons (Shadlen and Newsome, 1998) thus increasing the information that can be extracted from the IT population on a single trial (see Experimental Procedures for details).

Figure 2b shows the average performance as a function of population size for classifiers built to determine identity irrespective of position (blue line; "position-invariant" recognition), and for classifiers built to determine identity at particular positions (red line; "position-specific" recognition). Performance in both cases scales at a very similar rate as the population size grows. Notably, the absolute performance levels are very high for population sizes that are similar to those postulated to support visual discrimination tasks in other visual areas (Shadlen et al., 1996; >80% correct for a population of several hundred neurons). Figure 2c shows a detailed breakdown of performance across individual classification problems for a reasonably sized simulated population (10 simulated neurons for each recorded neuron; 680 simulated neurons total). For both

the position-invariant task (blue bars) and the position-specific task (red bars), each object was approximately equally well detected. Consistent with the fovea bias of IT neuronal RFs (Op de Beeck and Vogels, 2000), a slight improvement in object identification performance was found for the position-specific task at the center of gaze position (0 deg).

Discussion

In sum, these results show that even a small IT population simultaneously represents the identity and position of multiple objects in a format that is directly accessible to downstream neurons (up to three objects and three positions tested here). More generally, these results show that the same IT population is capable of *directly* supporting several tasks: position-invariant identification of isolated objects, position-invariant identification of objects when multiple objects are present, and object identification at each position when multiple objects are present.

At a quantitative level, the performance in each task as a function of population size shows that highly reliable performance on these tasks is achieved for population sizes that are similar to those postulated to support visual discrimination tasks in other visual areas (Shadlen et al., 1996). Indeed, this level of performance is likely a lower bound since we did explicitly search for selective neurons, and the full IT population likely contains a greater diversity of response profiles, rather than duplications of the same response profiles, which would tend to increase the amount of information available in the population (although correlated noise in neuronal responses can limit the effect of increasing population size; Zohary et al., 1994).

We have shown here that simultaneous readout of several object identities can be achieved with only brief image presentations (~100 ms) and simple weighted sums of mean firing rates over a short time interval (100 ms) across a modestly-sized population of IT neurons under passive viewing conditions. Our results cannot rule out a role for serial attention, synchronous firing, or other special mechanisms in disambiguating several, simultaneously present objects, nor do these results prove that the IT information we have exploited is actually used by downstream neurons. However, these results show that, without invoking any special mechanism, the IT population itself contains enough information to represent multiple object identities and positions in parallel (i.e. without incurring catastrophic representational interference).

While we show here that the IT population reliably represents several objects near the fovea in parallel, previous work suggests that attentional mechanisms can enhance such representations (e.g. Treisman and Schmidt, 1982). Furthermore, psychophysical studies (e.g. Posner et al., 1980) and IT physiological studies (Sheinberg and Logothetis, 2001; Rolls et al., 2003) suggest that attention and eye movements are required to maintain reliable IT representations for more eccentric objects, especially as scene complexity increases (e.g., as the number of objects increases).

More generally, our results are a reminder that even simple rate codes in populations of neurons can convey information that is not readily apparent from the responses of single units (Kohn and Movshon, 2004; Riesenhuber and Poggio, 1999). In particular, while examination of individual

IT neurons might suggest that their relative invariance to position (Gross et al., 1969; Ito et al., 1995) and relative lack of invariance to clutter (Miller et al., 1993; Missal et al., 1999; Rolls and Tovee, 1995; Zoccolan et al., 2005) might pose a problem in the representation of multiple objects, direct examination of population responses shows that no such problem exists. Interestingly, our results show that it is possible to read out the identity and position of each object at a level of performance that is comparable to the read-out of identity invariant to position (see Fig. 2). This finding indicates that this extra position-specific information is available at no extra "cost" relative to position-invariant recognition (i.e. the task that IT is traditionally thought to support). Thus, the incomplete position and clutter invariance observed in individual IT neurons may not represent a representational failure or shortcoming. Instead, it may reflect a balance at the population-level to support multiple task goals, including the ability to extract object identity despite image variation (e.g., due to position and object clutter), as well as the ability to simultaneously represent the identity and position of multiple objects. The factors and constraints that enter into such a balance could yield important insights into the neural coding of object information in IT.

Methods

Single Unit Recording. We recorded from 68 well-isolated neurons in anterior IT in two rhesus macaque monkeys (35 cells in monkey 1 and 33 in monkey 2). Surgical procedures, eye monitoring, and recording methods were done using standard techniques (DiCarlo and Maunsell, 2000), and were performed in accordance with the MIT Committee on Animal Care.

Stimuli and Tasks. Visual stimulus displays consisted of combinations of three possible object forms (star, triangle and cross shapes; white 57 Cd/m^2 on a gray background of 27 Cd/m^2) that could appear in three possible locations (at the center of gaze, 2º above, and 2º below, see Figure 1). All combinations of **a)** one object in each possible position (9 stimulus displays) **b)** two objects (without duplicates, 18 stimulus displays), and **c)** three objects (with no object repeated in the same display, 6 stimulus displays) were presented to the passively viewing monkey (33 stimulus displays total). On each behavioral trial, monkeys were required to fixate a central point while five stimulus displays were presented in pseudorandom order. Each stimulus display was presented for 100 ms followed by an inter-stimulus interval of 100 ms. This rate of five stimulus displays per second is roughly comparable to the timing of spontaneously generated saccades during recognition tasks (DiCarlo and Maunsell, 2000), and well within the timeframe that allows accurate object recognition (Potter, 1976). Both monkeys had been previously trained to perform an identification task with the three objects appearing randomly interleaved in each of the three positions (in isolation), and both monkeys achieved greater than 90% accuracy in this task. Monkeys performed this identification task while we advanced the electrode, and all isolated neurons that were responsive during this task (t-test; $p < 0.05$) were further studied with the 33 stimulus displays under the fixation conditions described above. Between 10 and 30 repetitions of each stimulus display were presented while recording from each IT neuron.

Neuronal responses. For each neuron, we computed spike counts over the time window from 100 to 200 ms post-stimulus onset for each presentation of each stimulus display (33 conditions total). The start of this time window was based on the well-known latency of IT neurons (Baylis

and Rolls, 1987). The end of the window is well below the reaction times of the monkeys when performing an identification task with these objects (DiCarlo and Maunsell, 2000), and is thus consistent with an integration window that could, in principle, be used by downstream neurons to extract object information.

Neuronal populations. By using the neuronal data to create synthesized IT neuronal populations with the same response profiles, we were able to explore the ability of a range of reasonably sized IT populations to support the various recognition tasks. To synthesize such populations, mean spike counts were computed for each of the 33 stimulus displays (described above), and these mean rates were used as the λ (rate) parameter of a Poisson random number generator to generate synthetic spike counts for 10 repetitions of each stimulus display. The spike counts of cortical neurons have been previously demonstrated in numerous contexts to be well approximated by Poisson statistics (Softky and Koch, 1992), and the response counts of the neuronal population measured here have a variance / mean ratio (Fano factor) close to unity (mean = 1.07), consistent with Poisson statistics. In all synthesized populations larger than the original set of recorded neurons (e.g. Figure 2b & c), multiple copies of each neuron were generated from the response profiles of each of the original recorded neurons. Because this procedure only uses response profiles from the recorded neurons, it cannot create stimulus selectivity that was not empirically observed, but it effectively minimizes the high trial-to-trial variability of cortical neurons in a manner that could be accomplished by downstream neurons integrating over larger population sizes (Shadlen and Newsome, 1998).

Performance shown in Figure 2c is based on a synthesized population with ten duplicates of each the original 68 neurons (i.e. a population of 680 neurons). Comparable performance levels could be achieved with smaller populations by only including neurons that showed significant selectivity for the objects in isolation at a level of $p < 0.05$ (1-way ANOVA, 29 neurons). The 290-neuron population obtained in this way achieved 78.8% mean accuracy in the position-invariant classification problems (blue bars in Figure 2), and 85.1% mean accuracy in the position-specific classification problems (red bars in Figure 2). In both cases this represents a performance reduction of less than 5% with the smaller population. Other, more complicated neuron selection procedures (e.g. not including neurons that provide redundant information) could possibly reduce the population size even further while maintaining high levels of performance.

Building linear discriminant classifiers. Across each IT population data set (actual or synthesized), the spike counts from each presentation of each stimulus display were assembled into *t* vectors of length *n*, where *t* is the number of trials used in the analysis (10 repetitions x 33 conditions), and *n* is the number of neurons. We analyzed this IT population response data set using Fisher linear discriminant analysis (Fisher, 1936). In its basic form, this technique takes labeled multivariate data points (here, *t* points in a *n*-dimensional space) belonging to two classes (e.g. "star present in top position" and "star not present in top position") and finds a hyperplane decision boundary that best separates the classes by maximizing the ratio of the difference between class means to the within-class variance. In practice, this involves first normalizing the input variables to have an identity within-class covariance matrix, and then finding the eigenvector corresponding to the largest eigenvalue of the between-class covariance matrix. A detailed discussion of linear discriminants can be found in Duda, Hart and Stork (2001). Although variants

on linear discriminant classification exist that can take into account the prior probability of a given label occurring, we explicitly did not use such variants to avoid improved guessing performance based purely on knowledge of the distribution of labels in the test set. We determined classification performance with respect to each particular question (e.g. Fig. 1) from each IT population data set using a leave-one-out cross-validation method in which one datum (i.e. one of the *t* points) is removed from the data set, the rest of the data is use to obtain the classifier boundary, and the removed point is tested for accurate classification. This operation is repeated for each of the *t* data and the overall performance is taken as the percentage of correct classifications across all of these tests. Standard errors of the mean performance (as in Figure 2) were computed by bootstrap resampling of these *t* cross-validation tests.

Since neurons were not recorded simultaneously, any trial-by-trial covariance structure between the dimensions of these vectors (i.e. between neurons) is meaningless and could lead to over-fitting and reduced classifier performance. To remove these spurious covariances (i.e. to "whiten" the covariance matrix), many data vectors were generated by randomly drawing which trials from each neuron went together into each vector (i.e. any stimulus display repetition from one neuron could "go with" any particular stimulus display repetition from another neuron), and repeating this procedure 10 times. Importantly, this procedure was performed *after* one to-be-classified data vector was removed for cross validation, maintaining the absolute independence of the *test* and *training* sets (i.e. there was no data in common between the *test* and *training* sets).

Object identification tasks. All problems were framed as simple two-class classification problems, in which a single linear discriminant (described above) was asked to report if a particular object was present or not, either in any position (blue bars in Figure 2), or just in one position, ignoring the presence or absence of objects in other positions (red bars in Figure 2). Within this formulation, classification consists simply of taking a weighted sum of input responses, and applying a threshold. As such, one could think of each classifier (and each bar in Figure 2c) as corresponding to a hypothetical downstream neuron, either hard-wired to perform a particular weighted sum, or perhaps dynamically set to this weighting by modulatory mechanisms. High performance across conditions where the target could appear in any position (blue bars in Figure 2a & c) and where only one position was relevant (red bars in Figure 2a & c) indicates that the same exact IT population can support the identity-extraction problem over a range of differently sized spatial regions.

Chance Performance. Because all classification problems undertaken in this study were two-class problems, "chance" (i.e. guessing) performance was 50% in all cases. Significance values were computed with respect to this chance level using the cumulative distribution function of a binomial distribution, which dictates analytically the probability of attaining empirically-observed performance levels, or greater, by random guessing. Given the large numbers of trials that go into each classification problem (10 repetitions x 33 stimulus displays) and the high levels of performance overall, the p-values associated with classifier performance often became infinitesimally small (i.e. it is extremely unlikely to achieve such high performance by guessing). Values well below $p = 10^{-6}$ are simply reported as $p << 10^{-6}$.

Since performance in determining which object was present in the center position was high, and because IT neurons typically respond best to objects at the center of gaze[8], we wondered how

chance levels for the non-central positions would change if we simply took the identity of the central object as given. Since we did not repeat the same object twice in one display, knowing that a given object was present in the central position would indicate that that object was not in the top or bottom position. These assumptions lead to a slight elevation of the chance level to 60.7%. However, the observed performances of 85.2% and 87.3% in the top and bottom positions, respectively (Figure 2c) are well above this level ($p << 10^{-6}$), indicating that the population responses contain highly significant information about which objects are present in all three locations.

Chapter 5:
Why is "Natural" Object Recognition Hard?

The construction of artificial vision systems and the study of biological vision are naturally intertwined, representing simultaneous efforts to forward and reverse engineer systems with similar goals. In recent years, large databases of "natural" images have become popular in the study of both biological and artificial vision, and some in the machine vision community have claimed impressive recognition performance using such image sets. However, we demonstrate here the danger inherent in the use of such image sets, showing that a simple V1-like model can outperform existing state-of-the-art object recognition systems on a popular standard 101-category image database, while at the same time failing on a "simpler" two-category image set specially constructed to better span the range of variation observed in the real world. In addition to tempering some of the claims of progress from the machine vision community, these results highlight the importance of real world variation and the difficulties in using "natural" images. We hope that these results ultimately point to new paths forward in the study of vision.

Visual object recognition is an extremely difficult computational problem. Any given object in the world can cast an essentially infinite number of different 2D images onto the retina as the object's position, size, pose, and lighting vary relative to the viewer. A recognition system, whether biological or artificial, must be able to extract the object's identity in spite of this variation. Artificial object recognition approaches seek to instantiate such recognition abilities in a computational system, sometimes with biological inspiration, sometimes without (Serre et al 2007, Lowe 2004, Zhang et al. 2006, Weber et al. 2000). Such computational approaches are critically important to understanding how the brain solves object recognition, because they can provide experimentally testable hypotheses and because instantiation of a working recognition system represents a particularly effective measure of success in understanding object recognition.

A major challenge in both biological and artificial object recognition is assessing performance – in part due to poor definition of what the recognition problem is. Ideally, artificial systems should be able to do what our own visual systems can. In practice, available computational power limits the number of images that can be tested, and the creation of large, labeled image exemplars can be extremely labor intensive. Because we are still at an early stage of understanding, we would like to tackle problems that provide "evolutionary" force, in that they are challenging in the right way, and can be made progressively more difficult. Partial success with a carefully constructed "easy" problem can often lead to more insight than complete failure on a problem that is far too difficult.

In recent years, "natural" images have become popular in the study of both biological and artifi-

cial vision (Gallant et al. 1998, Felsen and Dan 2005, Reinagel 2001, Bell and Sejnowski 1997, Simoncelli and Olshausen 2001). In artificial vision, the Caltech101 image set has a emerged as a "gold standard" for object recognition performance assessment (Li et al. 2004; http://www.vision.caltech.edu/Image_Datasets/Caltech101). The set consists of 101 classes of objects (e.g. planes, cars, faces, flamingos, etc. see Figure 1a) plus an additional "background" category (for 102 categories total). While a number of specific concerns have been raised with this set (see Ponce et al 2006 for more details), it is still widely used. The logic of the Caltech101 set (and sets like it; e.g. Caltech256, http://www.vision.caltech.edu/Image_Datasets/Caltech256) is that the sheer number of categories and the diversity of those images place a high bar for object recognition systems and require them to solve the core computational crux of the recognition problem. Because there are more than 100 categories, theoretical chance performance should be less than 1%. In recent years, several groups have reported what appears to be impressively high performance on this test – better than 60% correct across 102 categories (Wang et al. 2006, Mutch and Lowe 2006, Zhang et al. 2006, Lazbnik et al. 2006, Grauman and Darrell 2006).

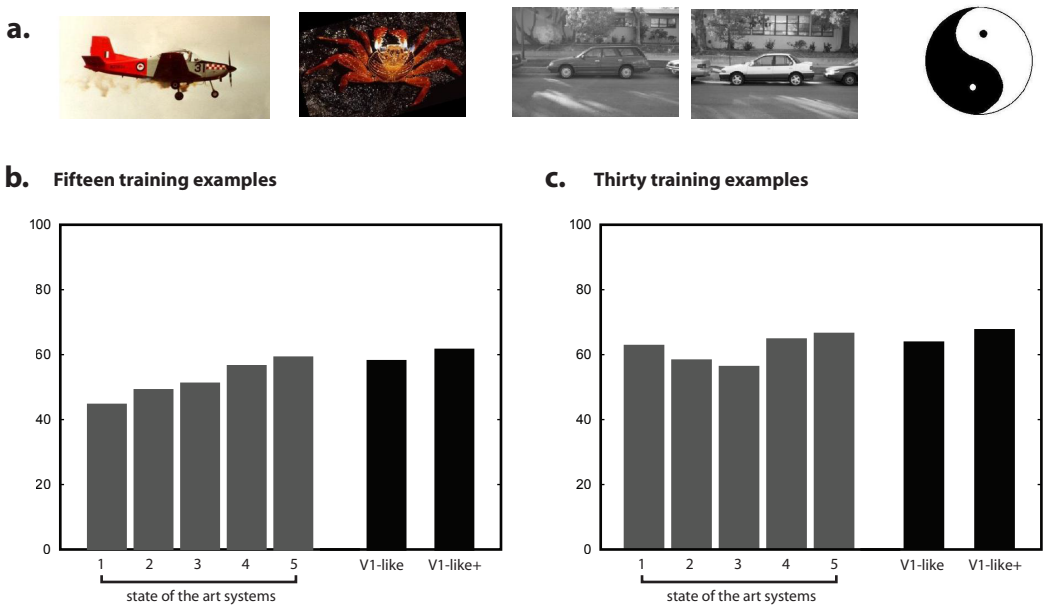


Figure 1. Performance of a simple V1-like model relative to current "state-of-the-art" objects recognition approaches on a standard image set. **a)** Examples images from the Caltech101 set. **b)** Performance results from five computational object recognition systems on the standard Caltech101 image set are shown in gray ([1] Wang et al. 2006, [2] Grauman and Darrell 2005, [3] Mutch and Lowe 2006, [4] Lazebnik et al. 2006, [5] Zhang et al. 2006). In this panel, fifteen training examples were used to train each system. Since chance performance on this 100+-way task should be less than 1%, performance values of greater than 40-50% have been taken as signs of great progress. The performance of our simple V1-like model, with and without additional "ad hoc" features (see Supplementary Methods) is shown in black. In spite of the fact that these models are extremely simple and lack any invariance-building mechanisms, they do as well as, or better than, the other object recognition systems from the literature. **c)** This panel shows the same performance values, but when thirty training examples were used, instead of fifteen.

However, it is not clear to what extent the Caltech101 test engages the core problem of object recognition. While the set certainly contains a large number of images (9144 images), the variation between and within classes is not controlled, and object backgrounds strongly covary with object category. The majority of images are also "composed" photographs, that is, a human decided how the shot should be framed, and thus the placement of objects within the image is not random and the set does not properly reflect the variation found in the real world. Furthermore, if the Caltech101 object recognition task is hard, it is not easy to know what makes it is hard – different kinds of variation (view, lighting, exemplar, etc.) are all inextricably mixed together. This is not just a problem with the Caltech101 set, it is also a problem that is endemic to other uncontrolled "natural" image sets (e.g. www.pascal-network.org/challenges/VOC).

To explore this issue, we constructed a very basic V1 simple cell-like recognition "model" (see Supplementary Methods for details) and tested it on the Caltech101 object recognition task, using the standard procedures published in the literature (Grauman and Darrell, 2006). To a neuroscientist, this model is a "null" model – it is arguably the simplest, most obvious starting point for describing the visual system. Importantly, it contains no mechanisms that could produce invariance, nor does it contain a particularly sophisticated representation of shape. It is a "strawman" model, and it should not be good for performing real-world object recognition tasks.

However, this simple V1-like model not only performs well on the "gold-standard" Caltech101 task, it outperforms all reported state-of-the-art computational efforts (Zhang et al 2006, Lazebnik et al 2006, Mutch and Lowe 2006, Wang et al 2006, Grauman and Darrell 2006) – our V1-like model achieves 61% and 67% correct with 15 and 30 training examples, respectively. Figure 1 shows the cross-validated performance of two versions of this simple model: one where only the model outputs (normalized, thresholded Gabor functions) are fed into a standard linear classifier, and one where some additional ad-hoc features are also used (e.g. local feature intensity histograms; see Supplementary Methods for details). In both cases, performance is comparable to, or better than, the current best reported performance in the literature. Portable code for building and evaluating this model is available online (http://web.mit.edu/dicarlo-lab/v1s/). Our claim from this result is not that our V1-like "model" is a good theory of recognition, or that Caltech101 and other such sets were not an important early step in establishing performance benchmarks. Instead, these results underscore that we must keep a clear understanding of why the problem is hard (Ullman 2000, Edelman 1999, Riesenhuber and Poggio 1999), that we must build performance tests that reflect that understanding, and that we should not assume that "natural" images automatically accomplish this goal.

To point a way forward, we constructed a series of "simpler" two-category image sets, consisting of rendered images of plane and car objects. By the logic of the Caltech101 test, this task should be substantially easier – there are only two object categories (rather than 101), and only a handful of specific objects per category (Figure 2a). In these sets, however, we explicitly and parametrically introduced real-world variation in the image that each object produced. In spite of the vastly smaller number of categories that the system was required to identify, the problem proved substantially harder for the V1-like "model", exactly as one would expect for an incomplete model of object recognition. Figure 2 shows how performance rapidly degrades toward chance-level as even modest amounts real-world object image variation are systematically intro-

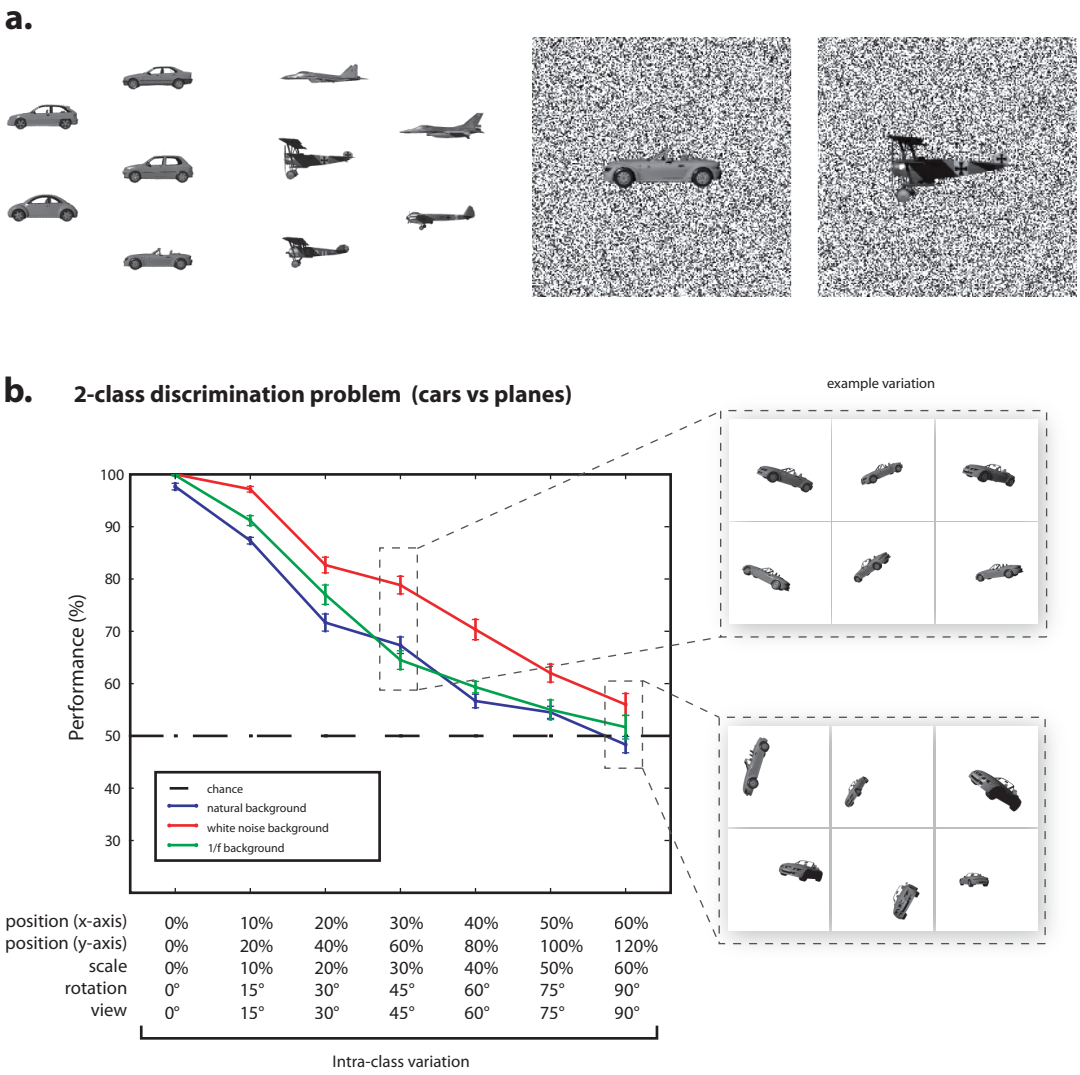


Figure 2. The same simple V1-like model that performed well in Figure 1, fails badly on a "simple" problem that requires tolerance to image variation. **a)** We used 3D models of cars and planes to generate a image sets for performing a cars-vs.-planes two-category test. By using 3D models, we were able to parametrically control the amount of view variation that the system was required to tolerate in order to perform the task. The models were rendered using raytracing software (POV-Ray), and were placed on a either a white noise background (shown here), a 1/f background, or a natural background (see Supplementary Methods). **b)** As the amount of variation in view parameters was increased (x-axis), performance drops off, eventually reaching chance level (50%). This result highlights a fundamental disconnect in the way object recognition sets are tested. By the logic of the Caltech101 set, this task should be easy, because it has so few categories (just two categories, as compared to the 100+ in the Caltech101). However, this V1-like model fails badly with this "easy" set, in spite of high performance with the supposedly more difficult "natural" image set.

duced. This result emphasizes that object recognition is hard, not because images are "natural" or "complex", but because each object can produce a very wide range of retinal images. Indeed this complexity can be a double-edged sword: although the addition of complex, natural backgrounds can make the problem more challenging, it can also make the problem easier (e.g. if the backgrounds highly covary with the object identity).

We argue that these results suggest that a new direction is needed to guide the development of object recognition systems. The issues cut deeper than simple performance evaluation – this is a question of how we think about the problem of object recognition and why it is hard. Large uncontrolled "natural" image sets may, on their face, seem to provide the best way to test real-world performance. However, as shown above, this is far from guaranteed. This question is also not simply an academic concern – great effort is now being expended to test models against a new, larger object recognition image sets (as if the smaller set has been solved), the "Caltech256." However, as with its predecessor, this new set fails to reflect real-world variation, and our "null" V1 model also performs well above chance (24% accuracy with 15 training examples to discriminate 257 categories, chance is 0.39%), and competitive with early published performance estimates on this new set (see Supplementary Figure 2).

How should we test progress in object recognition? One approach would be to a generate very large database of "natural" images, like the Caltech set, but captured in an unbiased way (i.e. with great care taken to avoid the implicit biases that occur in framing a snapshot). Done correctly, this approach has the advantage of directly representing the true problem domain. However, annotating such an image set is extremely labor intensive (but see the LabelMe project, Russell et al 2005). More importantly, a set that truly reflects all real-world variation may not provide evolutionary force to guide improvement in recognition models. That is, if the problem is too hard, it is not easy to construct a reduced version that still engages the core problem of object recognition.

Another approach, an extension of the one taken here, would be to use synthetic images, where ground truth can be known by design. In addition to obviating labor-intensive and error-prone labeling procedures, such an approach has the advantage that it can be parametrically made more difficult as needed (e.g. when a given model has achieved the ability to tolerate a certain amount of variation, a new instantiation of the test set with greater variation can be generated).

Supplementary Methods

A V1-like recognition system

Area V1 is the first stage of cortical processing of visual information and is the gateway of subsequent processing stages. We built a very basic representation inspired by known properties of V1 "simple" cells (a subpopulation of V1 cells). The responses of these cells to visual stimuli are well described by a linear filter, resembling a Gabor wavelet (<<ref hubel & wiesel >>), with a nonlinear output function (threshold and saturation) and some local normalization (analogous to "contrast gain control").

Operationally, our V1-like system consisted of the following processing steps:

Image pre-processing. First we converted the input image to grayscale and resized by bicubic interpolation the largest edge to a fixed size (150 pixels for Caltech datasets) while preserving its aspect ratio. The mean was subtracted from the resulting two-dimensional matrix and we divided it by its standard deviation (sphering). The resulting matrix had zero mean, unit variance and a size of HxW.

Local input normalization. For each pixel in the input matrix we subtracted the mean of a fixed window containing the pixel and its neighbor and we divided by the euclidean norm of the resulting vector if above a given threshold.

Linear filtering with a set of Gabor filters. We convolve the normalized image with a set of N two-dimensional Gabor filters with a fixed size, 16 orientations equally spaced around the clock and 6 spatial frequencies (1/2, 1/3, 1/4, 1/6, 1/11, 1/18) for a total of N=96 filters. Each filter has zero-mean and euclidean norm of one. The result is a three-dimensional matrix of size HxWxN where each two-dimensional slice is the output of each filter. To speed this step, the Gabor filters were decomposed via singular value decomposition into a form suitable for use in a separable convolution (this is possible because the Gabor filters are of low rank). The decomposed filters were constructed retain at least 90% of the variation in the filter.

Non-linear activation. The output of each gabor is passed through a standard output non-linearity corresponding to activation threshold and response saturation. Specifically all negative values of the three-dimensional matrix are set to 0 and all values greater than 1 are set to 1.

Output feature normalization. Finally, the output images (one per Gabor filter) were once again locally normalized as the inputs had been.

Classification

To test the utility of our V1-like representation for performing object recognition tasks, we performed a standard cross-validated classification procedure on the high-dimensional outputs of the model.

Dimensionality Reduction. To speed computation and improve classification performance, we reduced the dimensionality of the model output prior to classification. The output of this previous step was a stack of filtered images, one per Gabor filter. Because the dimensionality of this stack can be very high (up to 2,160,000 dimensions), standard dimensionality reduction techniques were used to prepare the data for classification. The output image stacks were low-pass filtered and down-sampled to a smaller size (30 by 30), such that the dimensionality was reduced to 86,400. The resulting data dimensionality was further reduced by PCA, keeping as many dimensions as there were data points in the training set. For the Caltech101 experiments (e.g. Figure 1) this dimensionality was 1530.

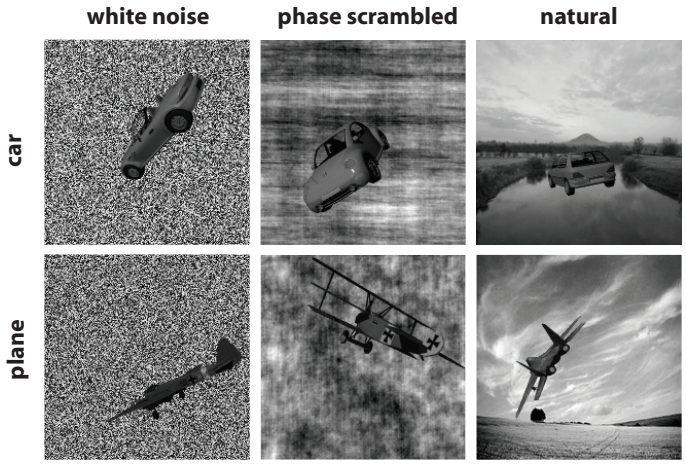
Additional "Ad Hoc" Features. To further explore the utility of this V1-like representation, we generated some additional easy-to-obtain features from our representation and concatenated these to the final feature vector, prior to PCA dimensionality reduction. These features included: raw grayscale input images (downsampled to 100x100 by bicubic interpolation; 10,000 features), color histograms (255 bins per color; 765 features), and local model output histograms (one per quadrant of the image) for each intermediate stage of the model: pre-normalization, post-normalization, and post downsampling (roughly 30,000 features total).

Throughout the text, results from the system containing these extra "ad hoc" features are reported separately from those obtained with the system that did not have these extra features. These extra features were added to demonstrate what was possible using additional obvious, "cheap" (but still fair) tricks that improve performance without incurring additional conceptual complexity.

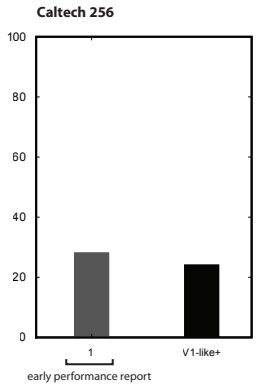
Training. Training and test images were carefully separated to ensure proper cross-validation. Fifteen training examples, and thirty testing examples were drawn from the full image set. Sphering parameters and PCA eigenvectors were computed from the training data, and the dimensionality-reduced training data were used to train a linear support vector machine (SVM) using libsvm-2.82 (Chang and Lin 2001). A one-versus-all approach was used to generate the multi-class SVM classifier. The testing data were then sphered using parameters determined from the training data and were projected onto the eigenvectors computed from the training data. The trained SVM was then used to classify each of the testing examples.

Testing protocol. Training and testing of the classifier uses a fixed number of examples when possible. The score reported is the average performance from 10 random splits of training and testing sets. Fifteen test images were classified per category, except in categories where there were not enough images available (in which case the minimum available was used). Since the Caltech 101 contains a different number of images for each category, care must be taken to ensure that per-category performance was normalized by the number of test examples considered in each category. This is a particular problem for the Caltech 101 set, because some of the largest categories are also empirically the easiest. For the performance values reported in this paper, categories were normalized such that the contribution of each category was equivalent. Results obtained with just 15 training examples and 15 testing examples (such that no normalization is required, because all categories have enough images) were appreciably similar to the results reported here.

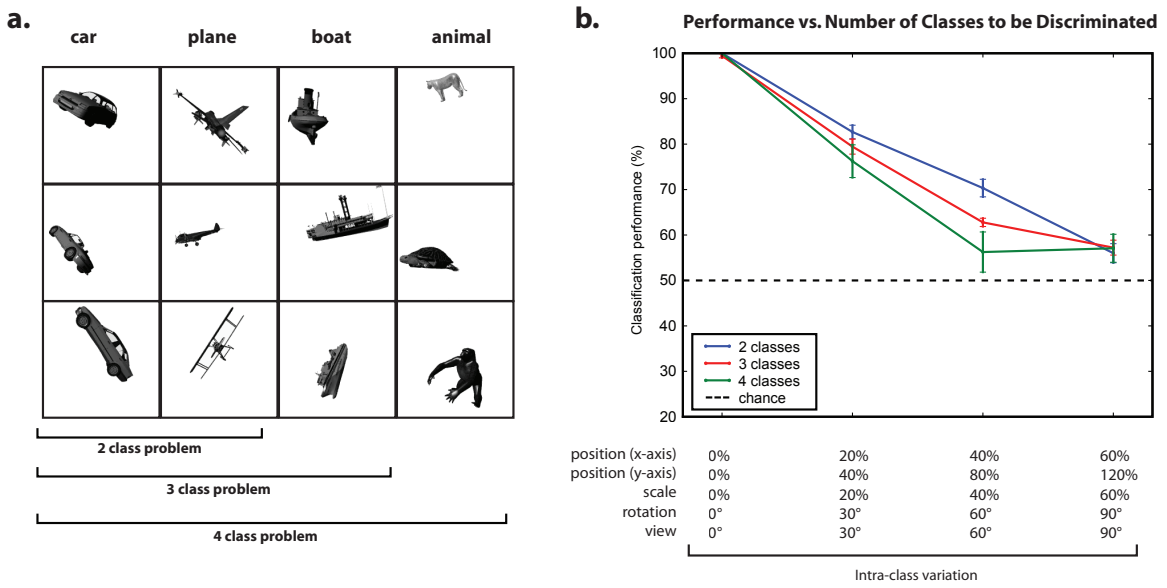
Supplementary Figures



Supplemental Figure 1. Backgrounds used. Model performance for our "simple" two class image set was assessed with the 3D models rendered onto a variety of backgrounds – white noise, phase-scrambled scene images, and intact scene images. Performance with each of these types of background is shown in Figure 2.



Supplemental Figure 2. Performance on the Caltech 256. [1] Griffin, Holub, and Perona 2007.



Supplemental Figure 3. Performance fall-off for increasing numbers of object categories. We previously showed that relatively modest amounts of image transformation push the performance of our simple V1-like model down to chance. Here we show that this fall-off becomes steeper as more categories-to-be-discriminated are added. **a)** Four categories of objects (cars, planes, boats, and animals) were used to measure performance when 2, 3 or 4 categories are considered **b)** Average identification performance ("is object category X present or not") is plotted as a function of view variation and number of object categories to be discriminated. Chance performance is 50% for all three lines, because average one-vs-all performance is shown here, not n-way recognition performance (i.e. "which object is present"). These plots show that performance falls to chance faster as the system is required to deal with more object categories.

References

Chapter 1

Afraz, S., Kiani, R. & Esteky, H. (2006) Microstimulation of inferotemporal cortex influences face categorization. Nature 442: 692-695.

Ashbridge, E. & Perrett, D. I. in Perceptual Constancy (eds. Walsh, V. & Kulikowski, J.) 192-209 (Cambridge University Press, Cambridge, 1998).

Barlow, H. in The Cognitive Neurosciences (ed. Gazzaniga).

Booth, M. C. A. & Rolls, E. T. (1998) View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. Cerebral Cortex 8, 510-523.

Brincat, S. L. & Connor, C. E. (2004) Underlying principles of visual shape selectivity in posterior inferotemporal cortex. Nat Neurosci 7, 880-6.

Chelazzi, L., Miller, E. K., Duncan, J. & Desimone, R. (1993) A neural basis for visual search in inferior temporal cortex. Nature 363, 345-347.

Connor, C. E., Gallant, J. L., Preddie, D. C. & Van Essen, D. C. (1996) Responses in area V4 depend on the spatial relationship between stimulus and attention. J Neurophysiol 75, 1306-8.

Cox, D. D., Meier, P., Oertelt, N. & DiCarlo, J. J. (2005) 'Breaking' position-invariant object recognition. Nat Neurosci 8, 1145-7.

David, S. V., Hayden, B. Y. & Gallant, J. L. (2006) Spectral receptive field properties explain shape selectivity in area V4. J Neurophysiol 96, 3492-505.

Desimone, R., Albright, T. D., Gross, C. G. & Bruce, C. (1984) Stimulus-selective properties of inferior temporal neurons in the macaque. J Neurosci 4, 2051-62.

DiCarlo, J. J. & Maunsell, J. H. R. (2000) Form representation in monkey inferotemporal cortex is virtually unaltered by free viewing. Nat Neurosci 3, 814-821.

DiCarlo, J. J. & Maunsell, J. H. R. (2003) Anterior Inferotemporal Neurons of Monkeys Engaged in Object Recognition Can be Highly Sensitive to Object Retinal Position. J Neurophysiol 89, 3264-78.

Duda, R. O., Hart, P. E. & Stork, D. G. (2001) Pattern Classification.

Edelman, S. Representation and Recognition in Vision (MIT Press, Cambridge, MA, 1999).

Edelman, S. & Intrator, N. (2002) Towards structural systematicity in distributed, statistically bound visual representations. 1-37.

Felleman, D. J. & Van Essen, D. C. (1991) Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex 1, 1-47.

Foldiak, P. (1991) Learning invariance from transformation sequences. Neural Computation 3, 194-200.

Gallant, J. L., Braun, J. & Van Essen, D. C. (1993) Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. Science 259, 100-3.

Gross, C. G. (2002) Genealogy of the "grandmother cell". Neuroscientist 8, 512-8.

Ghose, G. M. & Maunsell, J. H. (2002) Attentional modulation in visual cortex depends on task timing. Nature 419, 616-20.

Gross, C. G., Rocha-Miranda, C. E. & Bender, D. B. (1972) Visual properties of neurons in inferotemporal cortex of the Macaque. J Neurophysiol 35, 96-111.

Hubel, D. H. & Wiesel, T. N. (1977) Ferrier lecture. Functional architecture of macaque monkey visual cortex. Proc R Soc Lond B Biol Sci 198, 1-59.

Hung, C. P., Kreiman, G., Poggio, T. & DiCarlo, J. J. (2005) Fast readout of object identity from macaque inferior temporal cortex. Science 310, 863-6.

Ito, M., Tamura, H., Fujita, I. & Tanaka, K. (1995) Size and position invariance of neuronal responses in monkey inferotemporal cortex. Journal of Neurophysiology 73, 218-226.

Kayaert, G., Biederman, I. & Vogels, R. (2003) Shape tuning in macaque inferior temporal cortex. J Neurosci 23, 3016-27.

Kersten, D., Mamassian, P. & Yuille, A. (2004) Object perception as Bayesian inference. Annu Rev Psychol 55, 271-304.

Kreiman, G. et al. (2006) Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. Neuron 49, 433-45.

Logothetis, N. K., Pauls, J. & Poggio, T. (1995) Shape representation in the inferior temporal cortex of monkeys. Curr Biol 5, 552-63.

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

Logothetis, N. K. & Pauls, J. P. (1995) Psychophysical and physiological evidence for viewer-centered object representation in the primate. Cerebral Cortex 5, 270-288.

Logothetis, N. K. & Sheinberg, D. L. (1996) Visual object recognition. Ann. Rev. Neurosci. 19, 577-621.

Maunsell, J. H. R. (1995) The brain's visual world: representation of visual targets in cerebral cortex. Science 270, 764-9.

Mazer, J. A. & Gallant, J. L. (2003) Goal-related activity in V4 during free viewing visual search. Evidence for a ventral stream visual salience map. Neuron 40, 1241-50.

McAdams, C. J. & Maunsell, J. H. (2000) Attention to both space and feature modulates neuro-

nal responses in macaque area V4. J Neurophysiol 83, 1751-5.

McAdams, C. J. & Maunsell, J. H. (1999) Effects of attention on the reliability of individual neurons in monkey visual cortex [In Process Citation]. Neuron 23, 765-73.

Miller, E. K., Li, L. & Desimone, R. (1991) A neural mechanism for working and recognition memory in inferior temporal cortex. Science 254, 1377-1379.

Miyashita, Y. (1993) Inferior temporal cortex: where visual perception meets memory. Annual Review of Neuroscience 16, 245-263.

Motter, B. C. (1994) Neural correlates of attentive selection for color or luminance in extrastriate area V4. J Neurosci 14, 2178-89.

Moran, J. & Desimone, R. (1985) Selective attention gates visual processing in the extrastriate cortex. Science 229, 782-784.

Naya, Y., Yoshida, M., Takeda, M., Fujimichi, R. & Miyashita, Y. (2003) Delay-period activities in two subdivisions of monkey inferotemporal cortex during pair association memory task. Eur J Neurosci 18, 2915-8.

Op de Beeck, H. & Vogels, R. (2000) Spatial sensitivity of macaque inferior temporal neurons. J Comp Neurol 426, 505-18.

Olshausen, B. A. & Field, D. J. (2004) Sparse coding of sensory inputs. Curr Opin Neurobiol 14, 481-7.

Pasupathy, A. & Connor, C. E. (2001) Shape representation in area V4: position-specific tuning for boundary conformation. J Neurophysiol 86, 2505-19.

Perrett, D. I., Rolls, E. T. & Caan, W. (1982) Visual neurones responsive to faces in the monkey temporal cortex. Experimental Brain Research 47, 329-342.

Pollen, D. A., Przybyszewski, A. W., Rubin, M. A. & Foote, W. (2002) Spatial receptive field organization of macaque v4 neurons. Cereb Cortex 12, 601-16.

Potter, M. C. (1976) Short-term conceptual memory for pictures. J Exp Psychol [Hum Learn] 2, 509-22.

Reynolds, J. H. & Desimone, R. (2003) Interacting roles of attention and visual salience in V4. Neuron 37, 853-63.

Reynolds, J. H., Pasternak, T. & Desimone, R. (2000) Attention increases sensitivity of V4 neurons. Neuron 26, 703-14.

Riesenhuber, M. & Poggio, T. (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2, 1019-25.

Ringach, D. L. (2002) Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. J Neurophysiol 88, 455-63.

Rolls, E. T. (2000) Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. Neuron 27, 205-18.

Sary, G., Vogels, R. & Orban, G. A. (1993) Cue-invariant shape selectivity of macaque inferior temporal neurons. Science 260, 995-997.

Sato, T. (1988) Effects of attention and stimulus interaction on visual responses of inferior temporal neurons in macaque. J Neurophysiol 60, 344-64.

Schwartz, E. L., Desimone, R., Albright, T. D. & Gross, C. G. (1983) Shape recognition and inferior temporal neurons. Proceedings of the National Academy of Science (USA) 80, 5776-5778.

Sheinberg, D. L. & Logothetis, N. K. (2001) Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. J Neurosci 21, 1340-50.

Suzuki, W., Matsumoto, K. & Tanaka, K. (2006) Neuronal responses to object images in the macaque inferotemporal cortex at different stimulus discrimination levels. J Neurosci 26, 10524-35.

Tanaka, K. (1996) Inferotemporal cortex and object vision. Annual Review of Neuroscience 19, 109-139.

Tanaka, K. (2003) Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. Cereb Cortex 13, 90-9.

Tenenbaum JT, De Silva SV, Langford JC (2000). A global nonlinear framework for dimensionality reduction. Science 290(5500):2319-2323.

Tovée, M. J., Rolls, E. T. & Azzopardi, P. (1994) Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert monkey. Journal of Neurophysiology 72, 1049-1060.

Thorpe, S., Fize, D. & Marlot, C. (1996) Speed of processing in the human visual system. Nature 381, 520-522.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B. & Livingstone, M. S. (2006) A cortical region consisting entirely of face-selective cells. Science 311, 670-4.

Tsunoda, K. (2001) Complex objects are represented in macaque inferotemporal cortexby the combination of feature columns. Nat Neurosci 4, 832-838.

Ullman, S. High Level Vision (MIT Press, Cambridge, MA, 1996).

Ullman, S. & Soloviev, S. (1999) Computation of pattern invariance in brain-like structures. Neural Netw 12, 1021-1036.

Ungerleider, L. G. & Mishkin, M. in Analysis of Visual Behavior (eds. Ingle, D. J., Goodale, M. A. & Mansfield, R. J. W.) 549-585 (M.I.T. Press, Cambridge, MA, 1982).

Vogels, R. & Biederman, I. (2002) Effects of illumination intensity and direction on object coding in macaque inferior temporal cortex. Cereb Cortex 12, 756-66.

Vogels, R., Sáry, G. & Orban, G. A. (1995) How task-related are the responses of inferior temporal neurons? Visual Neuroscience 12, 207-214.

Wallis, G. & Rolls, E. T. (1997) Invariant face and object recognition in the visual system. Progress in Neurobiology 51, 167-194.

Wallis, G. & Bulthoff, H. H. (2001) Effects of temporal association on recognition memory. Proc Natl Acad Sci U S A 98, 4800-4.

Wandell, B. A. (1995) Foundations of Vision.

Wiskott, L. & Sejnowski, T. J. (2002) Slow feature analysis: unsupervised learning of invariances. Neural Comput 14, 715-70.

Yamane, Y., Tsunoda, K., Matsumoto, M., Phillips, A. N. & Tanifuji, M. (2006) Representation of the spatial relationship among object parts by neurons in macaque inferotemporal cortex. J Neurophysiol 96, 3147-56.

Chapter 2

Bedford, F. (1999) Keeping Perception Accurate. Trends in Cognitive Sciences 3, 4-12.

Biederman, I. & Bar, M. (1999) One-shot viewpoint invariance in matching novel objects. Vision Research 39, 2885-2899.

Dill, M. & Edelman, S. (2001) Imperfect invariance to object translation in the discrimination of complex shapes. Perception 30, 707-24.

Dill, M. & Fahle, M. (1998) Limited translation invariance of human pattern recognition. Perception & Psychophysics 60, 65-81.

Edelman, S. & Intrator, N. (2003) Towards structural systematicity in distributed, statically bound visual respresentations. Cognitive Science 27, 73-109.

Foldiak, P. (1991) Learning invariance from transformation sequences. Neural Computation 3, 194-200.

Karni, A. & Sagi, D. (1993) The time course of learning a visual skill. Nature 365, 250-2.

McConkie, G. W. & Currie, C. B. (1997) Visual stability across saccades while viewing complex pictures. J Exp Psychol Hum Percept Perform 22, 563-81 (1996).Wallis, G. & Rolls, E. T. Invariant face and object recognition in the visual system. Pro-gress in Neurobiology 51, 167-194.

Nazir, T. A. & O'Regan, J. K. (1990) Some results on translation invariance in the human visual system. Spat Vis 5, 81-100.

Ross, J., Morrone, M. C., Goldberg, M. E. & Burr, D. C. (2001) Changes in visual perception at the time of saccades. Trends Neurosci 24, 113-21.

Simoncelli, E. P. & Olshausen, B. A. (2001) Natural image statistics and neural representation. Annu Rev Neurosci 24, 1193-216.

Wallis, G. & Bülthoff, H. H. (2001) Effects of temporal association on recognition memory. Proc Natl Acad Sci U S A 98, 4800-4.

Watanabe, T., Nanez, J. E. & Sasaki, Y. (2001) Perceptual learning without perception. Nature 413, 844-8.

Wiskott, L. & Sejnowski, T. J. (2002) Slow feature analysis: unsupervised learning of invariances. Neural Comput 14, 715-70.

Chapter 3

Blanz V, Vetter T (1999) A morphable model for synthesis of 3D faces. In: 1999 Symposium on Interactive 3D Graphics - Proceedings of SIGGRAPH'99, pp 187-194. New York: ACM Press.

Britten KH, Heuer HW (1999) Spatial summation in the receptive fields of MT neurons. J Neurosci 19:5074-5084.

Carandini M, Heeger DJ, Movshon JA (1997) Linearity and normalization in simple cells of the macaque primary visual cortex. J Neurosci 17:8621-8644.

Cavanaugh JR, Bair W, Movshon JA (2002) Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. J Neurophysiol 88:2530-2546.

Chelazzi L, Duncan J, Miller EK, Desimone R (1998) Responses of neurons in inferior temporal cortex during memory-guided visual search. J Neurophysiol 80:2918-2940.

Connor CE, Preddie DC, Gallant JL, Van Essen DC (1997) Spatial attention effects in macaque area V4. J Neurosci 17:3201-3214.

Dean P (1976) Effects of inferotemporal lesions on the behavior of monkeys. Psychol Bull 83:41-71.

Dean P (1982) Visual behavior in monkeys with inferotemporal lesions. In: Analysis of Visual Behavior (Ingle D, Goodale M, Mansfield J, eds), pp 587-627: MIT Press.

Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. Annu Rev Neurosci 18:193-222.

DiCarlo JJ, Maunsell JH (2000) Form representation in monkey inferotemporal cortex is virtually unaltered by free viewing. Nat Neurosci 3:814-821.

DiCarlo JJ, Maunsell JH (2003) Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. J Neurophysiol 89:3264-3278.

Efron B, Tibshirani RJ (1998) An Introduction to the Bootstrap. Boca Raton, FL: Chapman & Hall/CRC.

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex 1:1-47.

Gawne TJ, Martin JM (2002) Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. J Neurophysiol 88:1128-1135.

Heeger DJ (1992) Normalization of cell responses in cat striate cortex. Vis Neurosci 9:181-197.

Heeger DJ, Simoncelli EP, Movshon JA (1996) Computational models of cortical visual processing. Proc Natl Acad Sci U S A 93:623-627.

Heuer HW, Britten KH (2002) Contrast dependence of response normalization in area MT of the rhesus macaque. J Neurophysiol 88:3398-3408.

Horel JA (1996) Perception, learning and identification studied with reversible suppression of cortical visual areas in monkeys. Behav Brain Res 76:199-214.

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

Keysers C, Xiao DK, Foldiak P, Perrett DI (2001) The speed of sight. J Cogn Neurosci 13:90-101.

Logothetis NK, Sheinberg DL (1996) Visual object recognition. Ann Rev Neurosci 19:577-621.

Logothetis NK, Pauls J, Poggio T (1995) Shape representation in the inferior temporal cortex of monkeys. Curr Biol 5:552-563.

Logothetis NK, Pauls J, Bulthoff HH, Poggio T (1994) View-dependent object recognition by monkeys. Curr Biol 4:401-414.

Maunsell JHR (1995) The brain's visual world: representation of visual targets in cerebral cortex. Science 270:764-769.

Miller EK, Gochin PM, Gross CG (1993) Suppression of visual responses of neurons in inferior temporal cortex of the awake macaque by addition of a second stimulus. Brain Res 616:25-29.

Missal M, Vogels R, Orban GA (1997) Responses of macaque inferior temporal neurons to overlapping shapes. Cereb Cortex 7:758-767.

Missal M, Vogels R, Li CY, Orban GA (1999) Shape interactions in macaque inferior temporal neurons. J Neurophysiol 82:131-142.

Moran J, Desimone R (1985) Selective attention gates visual processing in the extrastriate cortex. Science 229:782-784.

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

Poggio T, Bizzi E (2004) Generalization in vision and motor control. Nature 431:768-774.

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

Recanzone GH, Wurtz RH, Schwarz U (1997) Responses of MT and MST neurons to one and two moving objects in the receptive field. J Neurophysiol 78:2904-2915.

Reynolds JH, Desimone R (2003) Interacting roles of attention and visual salience in V4. Neuron 37:853-863.

Reynolds JH, Chelazzi L, Desimone R (1999) Competitive mechanisms subserve attention in macaque areas V2 and V4. J Neurosci 19:1736-1753.

Rice JA (1995) Mathematical Statistics and Data Analysis. Belmont, CA: Duxbury Press.

Rieke F, Warland D, Ruyter van Steveninck RR, Bialek W (1997) Spikes: exploring the neural code. Cambridge, MA: MIT Press.

Riesenhuber M, Poggio T (1999a) Hierarchical models of object recognition in cortex. Nat Neurosci 2:1019-1025.

Riesenhuber M, Poggio T (1999b) Are cortical models really bound by the "binding problem"? Neuron 24:87-93, 111-125.

Rogers DF (2000) An Introduction to NURBS with Historical Perspective. San Franscisco, CA: Morgan Kaufmann Publishers.

Rolls ET, Tovee MJ (1995) The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the receptive field. Exp Brain Res 103:409-420.

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

Rousselet GA, Thorpe SJ, Fabre-Thorpe M (2003) Taking the MAX from neuronal responses. Trends Cogn Sci 7:99-102.

Rousselet GA, Thorpe SJ, Fabre-Thorpe M (2004) How parallel is visual processing in the ventral pathway? Trends Cogn Sci 8:363-370.

Rubin GS, Turano K (1992) Reading without saccadic eye movements. Vision Res 32:895-902.

Sato T (1989) Interactions of visual stimuli in the receptive fields of inferior temporal neurons in awake macaques. Experimental Brain Research 77:23-30.

Schwartz O, Simoncelli EP (2001) Natural signal statistics and sensory gain control. Nat Neurosci 4:819-825.

Shadlen MN, Newsome WT (1994) Noise, neural codes and cortical organization. Curr Opin Neurobiol 4:569-579.

Shadlen MN, Newsome WT (1998) The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. J Neurosci 18:3870-3896.

Sheinberg DL, Logothetis NK (2001) Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. J Neurosci 21:1340-1350.

Shelton C (2000) Morphable surface models. Int J Comp Vis 38:75-91.

Softky WR, Koch C (1993) The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. J Neurosci 13:334-350.

Tanaka K (1996) Inferotemporal cortex and object vision. Annual Review of Neuroscience 19:109-139.

Treue S, Maunsell JH (1996) Attentional modulation of visual motion processing in cortical areas MT and MST. Nature 382:539-541.

Weiskrantz L, Saunders RC (1984) Impairments of visual object transforms in monkeys. Brain 107:1033-1072.

Chapter 4

Baylis, G. C., and Rolls, E. T. (1987). Responses of neurons in the inferior temporal cortex in short term and serial recognition memory tasks. Exp Brain Res 65, 614-622.

Braitenberg, V. (1978). Cortical Architectonics: General and Areal. In Architectonics of the Cerebral Cortex, M. a. P. H. Brazier, ed. (New York, Raven).

DiCarlo, J. J., and Maunsell, J. H. R. (2000). Form representation in monkey inferotemporal cortex is virtually unaltered by free viewing. Nat Neurosci 3, 814-821.

DiCarlo, J. J., and Maunsell, J. H. R. (2003). Anterior Inferotemporal Neurons of Monkeys Engaged in Object Recognition Can be Highly Sensitive to Object Retinal Position. J Neurophysiol 89, 3264-3278.

Duda, R. O., Hart, P.E., and Stork, D.G. (2001). Pattern Analysis (New York, Wiley-Interscience).

Edelman, S., and Intrator, N. (2003). Towards structural systematicity in distributed, statistically bound visual representations. Cognitive Science, 27:73-110.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179-188.

Gochin, P. M. (1994). Properties of simulated neurons from a model of primate inferior temporal cortex. Cereb Cortex 4, 532-543.

Gross, C. G., Bender, D. B., and Rocha-Miranda, C. E. (1969). Visual receptive fields of neurons in inferotemporal cortex of the monkey. Science 166, 1303-1307.

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

Kohn, A., and Movshon, J. A. (2004). Adaptation changes the direction tuning of macaque MT neurons. Nat Neurosci 7, 764-772.

Logothetis, N. K., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. Curr Biol 5, 552-563.

McCulloch, W. S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bull Math Biol 52, 99-115; discussion 173-197.

Miller, E. K., Gochin, P. M., and Gross, C. G. (1993). Suppression of visual responses of neurons in inferior temporal cortex of the awake macaque by addition of a second stimulus. Brain Res 616, 25-29.

Missal, M., Vogels, R., Li, C., and Orban, G. A. (1999). Shape interactions in macaque inferior temporal neurons. J Neurophysiol 82, 131-142.

Op de Beeck, H., and Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. J Comp Neurol 426, 505-518.

Posner, M. I., Snyder, C. R., and Davidson, B. J. (1980). Attention and the detection of signals. J Exp Psychol 109, 160-174.

Potter, M. C. (1976). Short-term conceptual memory for pictures. J Exp Psychol Hum Learn 2, 509-522.

Riesenhuber, M., and Poggio, T. (1999). Are cortical models really bound by the "binding problem?" Neuron 24, 87-93, 111-125.

Rolls, E. T., Aggelopoulos, N. C., and Zheng, F. (2003). The receptive fields of inferior temporal cortex neurons in natural scenes. J Neurosci 23, 339-348.

Rolls, E. T., and Tovee, M. J. (1995). The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the receptive field. Exp Brain Res 103, 409-420.

Rousselet, G. A., Thorpe, S. J., and Fabre-Thorpe, M. (2004). How parallel is visual processing in the ventral pathway? Trends Cogn Sci 8, 363-370.

Schwartz, E. L., Desimone, R., Albright, T. D., and Gross, C. G. (1983). Shape recognition and inferior temporal neurons. PNAS 80, 5776-5778.

Shadlen, M. N., Britten, K. H., Newsome, W. T., and Movshon, J. A. (1996). A computational analysis of the relationship between neuronal and behavioral responses to visual motion. J Neurosci 16, 1486-1510.

Shadlen, M. N., and Newsome, W. T. (1998). The variable discharge of cortical neurons: Implications for connectivity, computation and information coding. J Neurosci 18, 3870-3896.

Shannon, C. E. (1963). The mathematical theory of communication. MD Comput 14, 306-317.

Sheinberg, D. L., and Logothetis, N. K. (2001). Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. J Neurosci 21, 1340-1350.

Softky, W. R., and Koch, C. (1992). Cortical cells should fire regularly, but do not. Neural computation 4, 643-646.

Tovee, M. J., Rolls, E. T., Treves, A., and Bellis, R. P. (1993). Information encoding and the responses of single neurons in the primate temporal visual cortex. J Neurophysiol 70, 640-654.

Treisman, A. (1999). Solutions to the binding problem: progress through controversy and convergence. Neuron 24(1), 105-110, 111-125.

Treisman, A., and Schmidt, H. (1982). Illusory conjunctions in the perception of objects. Cognit Psychol 14, 107-141.

Ungerleider, L. G., and Mishkin, M. (1982). Two cortical visual systems. In Analysis of Visual Behavior, D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, eds. (Cambridge, MA, M.I.T. Press), pp. 549-585.

von der Malsburg, C. (1999). The what and why of binding: the modeler's perspective. Neuron 24, 95-104, 111-125.

Wainwright, M.J., Schwartz, O., Simoncelli, E.P. (2002). Natural Image Statistics and Divisive Normalization: Modeling Nonlinearity and Adaptation in Cortical Neurons. In Probabilistic Models of the Brain: Perception and Neural Function, Olshausen B, Rao R, Lewicki M, eds. (Cambridge, MIT Press).

Zoccolan, D., Cox, D. D., and DiCarlo, J. J. (2005). Multiple object response normalization in monkey inferotemporal cortex. J Neurosci 25, 8150-8164.

Zohary, E., Shadlen, M. N., and Newsome, W. T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. Nature 370, 140-143.

Chapter 5

Bell AJ, Sejowski TJ (1997) The "independent components" of natural scenes are edge filters. 37(23):3327-3338.

Edelman S (1999) Representation and recognition in vision. MIT Press.

Felsen G, Dan Y (2005) A natural approach to studying vision. Nature Neuroscience 8:1643-1646.

Gallant JL, Conor CE, Van Essen DC (1998) Neural activity in areas V1, V2, and V4 during free viewing of natural scenes compared to control images. Neuroreport 9(1):85-89.

Grauman K and Darrell T (2006) Pyramid match kernels: Discriminative classification with sets of image features. CSAIL Technical Report: CSAIL-TR-2006-20.

Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. IEEE CVPR 2006.

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2):91-110.

Mutch J, Lowe DG (2006) Multiclass object recognition with sparse, localized features. IEEE CVPR 2006 1:11-18.

Ponce J, Berg TL, Everingham H, Forsyth DA, Hebert M, Lazebnik S, Marszalek M, Schmid C, Russell BC, Torralba A, Williams CKI, Zhang J, Zisserman A (in press) Dataset Issues in Object Recognition. *Toward Category-Level Object Recognition*, Springer-Verlag Lecture Notes in Computer Science. J. Ponce, M. Hebert, C. Schmid, and A. Zisserman (eds.)

Reinagel P (2001) How do visual neurons respond in the real world. Current Opinion in Neurobiology. 11(4):437-442.

Russel B, Torralba A, Murphy K, Freeman WT (2005) LabelMe: a database and web-based tool for annotation. MIT AI Lab Memo AIM-2005-025.

Simoncelli EP, Olshausen BA (2001) Natural image statistics and neural representation. Annual Review of Neuroscience 24(1):1193-1216.

Ullman S (2000) High-level vision: object recognition and visual cognition. MIT Press.

Wang G, Zhang Y, Li FF (2006) Regions for object categorization in a generative framework. IEEE CVPR 20006 2:1597-1604.

Weber M, Welling M, Perona P (2000) Unsupervised learning of models for recognition. Proc. ECCV 1:18-32.

Zhang H, Berg A, Maire M, Malik, J (2006) SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. IEEE CVPR 2006 2:2126-2136.