massachusetts institute of technology — computer science and artificial intelligence laboratory

A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex

T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, T. Poggio

AI Memo 2005-036 December 2005
CBCL Memo 259

A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex

Thomas Serre, Minjoon Kouh, Charles Cadieu, Ulf Knoblich, Gabriel Kreiman and Tomaso Poggio[1]

Center for Biological and Computational Learning, McGovern Institute for Brain Research, Computer Science and Artificial Intelligence Laboratory, Brain Sciences Department, Massachusetts Institute of Technology

Abstract

We describe a quantitative theory to account for the computations performed by the feedforward path of the ventral stream of visual cortex and the local circuits implementing them. We show that a model instantiating the theory is capable of performing recognition on datasets of complex images at the level of human observers in rapid categorization tasks. We also show that the theory is consistent with (and in some case has predicted) several properties of neurons in V1, V4, IT and PFC. The theory seems sufficiently comprehensive, detailed and satisfactory to represent an interesting challenge for physiologists and modelers: either disprove its basic features or propose alternative theories of equivalent scope. The theory suggests a number of open questions for visual physiology and psychophysics.

This version replaces the preliminary "Halloween" CBCL paper from Nov. 2005.

This report describes research done within the Center for Biological & Computational Learning in the Department of Brain & Cognitive Sciences and in the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology.

This research was sponsored by grants from: Office of Naval Research (DARPA) under contract No. N00014-00-1-0907, National Science Foundation (ITR) under contract No. IIS-0085836, National Science Foundation (KDI) under contract No. DMS-9872936, and National Science Foundation under contract No. IIS-9800032

Additional support was provided by: Central Research Institute of Electric Power Industry, Center for e-Business (MIT), Eastman Kodak Company, DaimlerChrysler AG, Compaq, Honda R&D Co., Ltd., Komatsu Ltd., Merrill-Lynch, NEC Fund, Nippon Telegraph & Telephone, Siemens Corporate Research, Inc., The Whitaker Foundation, and the SLOAN Foundations.

[1]To whom correspondence should be addressed.

Contents

1 Introduction

Preface By now, there are probably several hundreds models about visual cortex. The very large majority deals with specific visual phenomena (such as specific visual illusions) or with specific cortical areas or specific circuits. Some of them have provided a useful contribution to Neuroscience and a few had an impact even on physiologists [Carandini and Heeger, 1994; Reynolds et al., 1999]. Very few address a generic, high-level computational function such as object recognition (see [Fukushima, 1980; Amit and Mascaro, 2003; Wersing and Koerner, 2003; Perrett and Oram, 1993]). We are not aware of any model which does it in a quantitative way while being consistent with psychophysical data on recognition and physiological data throughout the different areas of visual cortex while using plausible neural circuits. In this paper, we propose a quantitative theory of object recognition in primate visual cortex that 1) bridges several levels, from biophysics to physiology, to behavior and 2) achieves human level performance in rapid recognition of complex natural images. The theory is restricted to the feedforward path of the ventral stream and therefore to the first 150 ms or so of visual recognition; it does not describe top-down influences, though it is in principle capable of incorporating them.

Recognition is computationally difficult. The visual system rapidly and effortlessly recognizes a large number of diverse objects in cluttered, natural scenes. In particular, it can easily categorize images or parts of them, for instance as faces, and identify a specific one. Despite the ease with which we see, visual recognition – one of the key issues addressed in computer vision – is quite difficult for computers and is indeed widely acknowledged as a very difficult computational problem. The problem of object recognition is even more difficult from the point of view of Neuroscience, since it involves several levels of understanding from the information processing or computational level to the level of circuits and of cellular and biophysical mechanisms. After decades of work in striate and extrastriate cortical areas that have produced a significant and rapidly increasing amount of data, the emerging picture of how cortex performs object recognition is in fact becoming too complex for any simple, qualitative "mental" model. It is our belief that a quantitative, computational theory can provide a much needed framework for summarizing and organizing existing data and for planning, coordinating and interpreting new experiments.

Recognition is a difficult trade-off between selectivity and invariance. The key computational issue in object recognition is the specificity-invariance trade-off: recognition must be able to finely discriminate between different objects or object classes while at the same time be tolerant to object transformations such as scaling, translation, illumination, viewpoint changes, change in context and clutter, non-rigid transformations (such as a change of facial expression) and, for the case of categorization, also to shape variations within a class. Thus the main computational difficulty of object recognition is achieving a very good trade-off between selectivity and invariance.

Architecture and function of the ventral visual stream. Object recognition in cortex is thought to be mediated by the ventral visual pathway [Ungerleider and Haxby, 1994] running from primary visual cortex, V1, over extrastriate visual areas V2 and V4 to inferotemporal cortex, IT. Based on physiological experiments in monkeys, IT has been postulated to play a central role in object recognition. IT in turn is a major source of input to PFC involved in linking perception to memory and action [Miller, 2000].

Over the last decade, several physiological studies in non-human primates have established a core of basic facts about cortical mechanisms of recognition that seem to be widely accepted and that confirm and refine older data from neuropsychology. A brief summary of this consensus of knowledge begins with the groundbreaking work of Hubel & Wiesel first in the cat [Hubel and Wiesel, 1962, 1965b] and then in the macaque monkey [Hubel and Wiesel, 1968]. Starting from *simple cells* in primary visual cortex, V1, with small receptive fields that respond preferably to oriented bars, neurons along the ventral stream [Perrett and Oram, 1993; Tanaka, 1996; Logothetis and Sheinberg, 1996] show an increase in receptive field size as well as in the complexity of their preferred stimuli [Kobatake and Tanaka, 1994]. At the top of the ventral stream, in anterior inferotemporal cortex (AIT), cells are tuned to complex stimuli such as faces [Gross et al., 1972; Desimone et al., 1984; Desimone, 1991; Perrett et al., 1992].

The tuning of the view-tuned and object-tuned cells in AIT depends on visual experience as shown by [Logothetis et al., 1995] and supported by [Kobatake et al., 1998; DiCarlo and Maunsell, 2000; Logothetis et al., 1995; Booth and Rolls, 1998]. A hallmark of these IT cells is the robustness of their firing to stimulus transformations such as scale and position changes [Tanaka, 1996; Logothetis and Sheinberg, 1996; Logothetis et al., 1995; Perrett and Oram, 1993]. In addition, as other studies have shown [Perrett and Oram, 1993; Booth and Rolls, 1998; Logothetis et al., 1995; Hietanen et al., 1992], most neurons show specificity for a certain object view or lighting condition. In particular, Logothetis *et al.* [Logothetis et al., 1995] trained monkeys to perform an object recognition task with isolated views of novel 3D objects (paperclips, see [Logothetis et al., 1995]). When recording from the animals' IT, they found that the great majority of neurons selectively tuned to the training objects were view-tuned (with a half-width of about 20° for rotation in depth) to one of the training objects (about one tenth of the tuned neurons were view-invariant, in agreement with earlier predictions [Poggio and Edelman, 1990]), but exhibited an average translation invariance of 4° (for typical stimulus sizes of 2°) and an average scale invariance of two octaves [Riesenhuber and Poggio, 1999b]. Whereas view-invariant recognition requires visual experience of the specific novel object, significant position and scale invariance seems to be immediately present in the view-tuned neurons [Logothetis et al., 1995] without the need of visual experience for views *of the specific object* at different positions and scales (see also [Hung et al., 2005a]. Whether invariance to a particular transformation requires experience of the specific object or not may depend on the similarity of the different views as assessed by the need to access 3D information of the object (*e.g.,* for in-depth rotations) or incorporate properties about its material or reflectivity (*e.g.,* for changes in illumination), see Note 4.

In summary, the accumulated evidence points to four, mostly accepted, properties of the feedforward path of the ventral stream architecture: a) A hierarchical build-up of invariances first to position and scale (importantly, scale and position invariance — over a restricted range — do not require learning specific to an individual object) and then to viewpoint and other transformations (note that invariance to viewpoint, illumination etc. requires visual experience of several different views of the specific object); b) An increasing number of subunits, originating from inputs from previous layers and areas, with a parallel increase in size of the receptive fields and potential complexity of the optimal stimulus [1]; c) A basic feedforward processing of information (for "immediate" recognition tasks); d) Plasticity and learning probably at all stages with a time scale that decreases from V1 to IT and PFC.

A theory of the ventral stream After the breakthrough recordings in V1 by Hubel & Wiesel there has been a noticeable dearth of comprehensive theories attempting to explain the function and the architecture of visual cortex beyond V1. On the other hand myriads of specific models have been suggested to "explain" specific effects, such as contrast adaptation or specific visual illusions. The reason of course is that a comprehensive theory is much more difficult, since it is highly constrained by many different data from anatomy and physiology at different stages of the ventral stream and by the requirement of matching human performance in complex visual tasks such as object recognition.

We believe that computational ideas and experimental data are now making it possible to begin describing a satisfactory quantitative theory of the ventral stream focused on explaining visual recognition. The theory may well be incorrect – but at least it represents a skeleton set of claims and ideas that deserve to be either falsified or further developed and refined.

The theory described in this paper has evolved over the last 6 years from a model introduced in [Riesenhuber and Poggio, 1999b], as the result of computer simulations, new published data and especially collaborations and interactions with several experimental labs (Logothetis in the early years and now Ferster, Miller, DiCarlo, Lampl, Freiwald, Livingstone, Connor, Hegde and van Essen). The theory includes now passive learning to account for the tuning and invariance properties of neurons from V2 to IT. When exposed to many natural images the model generates a large set of shape-tuned units which can be interpreted as a universal (redundant) dictionary of shape-components with the properties of overcompleteness and non-uniqueness. When tested on real-world natural images, the model outperforms the best computer vision systems on several different recognition tasks. The model is also consistent with many – though not all – experimental data concerning the anatomy and the physiology of the main visual areas of cortex, from V1 to IT.

As required, the theory bridges several levels of understanding from the computational and psychophysical one to the level of system physiology and anatomy to the level of specific microcircuits and biophysical properties. Our approach is more definite at the level of the system computations and architecture. It is more tentative at the level of the biophysics, where we are limited to describing *plausible* circuits and mechanisms that could be used by the brain.

This version of the theory is restricted to the feedforward path in the ventral stream It is important to emphasize from the outset the basic assumption and the basic limitation of the current theory: we only consider the first 150 ms of the flow of information in the ventral stream — behaviorally equivalent to considering "immediate recognition" tasks — since we assume that this flow during this short period of time is likely to be mainly feedforward across visual areas (of course, anatomical work suggests that local connectivity is even more abundant than feedforward connectivity [Binzegger et al., 2004]; local feedback loops almost certainly have key roles, as they do in our theory, see later and see [Perrett and Oram, 1993]).

It is well known that recognition is possible for scenes viewed in rapid visual presentation that do not allow sufficient time for eye movements or shifts of attention [Potter, 1975]. Furthermore, EEG studies [Thorpe et al., 1996] provide evidence that the human visual system is able to solve an object detection task – determining whether a natural scene contained an animal or not – within 150 ms. Extensive evidence [Perrett et al., 1992] shows that the onset of the response in IT neurons begins 80-100 ms after onset of the visual stimulus and the response is tuned to the stimulus essentially from the very beginning [Keysers et al., 2001]. Recent data [Hung et al., 2005a] show that the activity of small neuronal populations (around 100 randomly selected cells) in IT over very short time intervals (as small as 12.5 ms) after beginning of the neural response (80-100 ms after onset of the stimulus) contains surprisingly accurate and robust information supporting a variety of recognition tasks. Finally, we know that the animal detection task [Thorpe et al., 1996] can be carried out without top-down attention [Li et al., 2002]. Again, we wish to emphasize that none of this rules out the use of local feedback – which is in fact used by the circuits we propose for the main two operations postulated by the theory (see Section 5) – but suggests a hierarchical forward architecture as the core architecture underlying "immediate" recognition.

Thus *we ignore any dynamics of the back-projections* and focus the paper on the feedforward architecture of the visual stream and its role in the first 150 ms or so of visual perception. The basic steps for rapid categorization are likely to be completed in this time, including tuned responses of neurons in IT. To be more precise, the theory assumes that back-projections may play a "priming" role in setting up "routines" in PFC or even earlier than IT – for simplicity let us think of "routines" as modulations of specific synaptic weights – in a task-dependent way before stimulus presentation but it also assumes that backprojections do not play a major dynamic role during the first 150 ms of recognition.

Normal vision and back-projections: a preliminary, qualitative framework Of course, a complete theory of vertebrate vision must take into account multiple fixations, image sequences, as well as top-down signals, attentional effects and the structures mediating them (*e.g.*, the extensive back-projections present throughout cortex). Thus, though our model at present ignores the effect of back-projections (or to be more precise it assumes that there is no change in their effects during the short time intervals we consider here), we state here our presently tentative framework for eventually incorporating their role.

The basic idea – which is not new and more or less accepted in these general terms – is that one key role of back-projections is to select and modulate specific connections in early areas in a top-down fashion – in addition to manage and control learning processes. Back-Projections may effectively run "programs" for reading out specific task-dependent information from IT (for instance, one program may correspond to the question "is the object in the scene an animal?", another may read out information about the size of the object in the image from activity in IT [Hung et al., 2005a]). They may also select "programs" in areas lower than IT (probably by modulating connection weights). During normal vision, back-projections are likely to control in a dynamic way routines running at all levels of the visual system throughout attentional shifts (and fixations). In particular, small areas of the visual fields may be "routed" from the appropriate early visual area (as early as V1) by covert attentional shift controlled from the top to circuits specialized for any number of specific tasks – such as vernier discrimination (see [Poggio and Edelman, 1990] and Poggio's AI Working Paper 258, "Routing Thoughts", 1984). The routing mechanism could achieve the desired invariances to position, scale and orientation and thereby reduce the complexity of learning the specific task.

This highly speculative framework fits best with the point of view described by [Hochstein and Ahissar, 2002]. Its emphasis is thus somewhat different with respect to ideas related to prediction-verification recursions – an approach known in AI as "hypothesis-verification" (see among others, [Hawkins and Blakeslee, 2002; Mumford, 1996; Rao and Ballard, 1999]). Hochstein and Ahissar suggested that explicit vision advances in reverse hierarchical direction, starting with "vision at a glance" (corresponding to our "immediate recognition") at the top of the cortical hierarchy and returning downward as needed in a "vision with scrutiny" mode in which reverse hierarchy routines focus attention to specific, active, low-level units. Of course, there is a large gap between all of these ideas and a quantitative theory of the back-projections such as the one described in this paper for the feedforward path in the ventral stream.

Plan of the paper The plan of this memo is as follows. We describe in the next section (Section 2) the theory, starting from its architecture, its two key operations and its learning stages. Section 3 shows that an implementation of the theory achieves good recognition results on natural images (compared with computer vision systems) and – more importantly – mimics human performance on rapid categorization tasks. We then review the evidence (section 4) about the agreement of the model with single cell recordings in visual cortical areas (V1, V2, V4, IT). In Section 5 we describe some of the possible "microcircuits" which may implement the key operations assumed by the theory and discuss a possible "canonical" microcircuit at their core. The final Section 6 discusses the state of the theory, its limitations, a number of open questions, including critical experiments, and its extension to include top-down effects and cortical back-projections. Throughout the paper, most of the details can be found in the appendices.

Notes

[1]The connection between complexity and size of the receptive field through the number of subunits follows in fact from our theory (see later). The subunits are of the V1 simple cell type and possibly also of the LGN center-surround type.

[2]In this paper, we use the term *categorization* to designate *between-class* object classification, the term *identification* for classification *within* an object class and the term *recognition* for either task. Computationally, there is no difference between categorization and identification (see [Riesenhuber and Poggio, 2000]).

[3]We have already used an earlier version of the theoretical framework described here – and its corresponding model simulations – in on-going collaborations with physiology labs to drive a highly multidisciplinary enterprise. Models provide a way to summarize and integrate the data, to check their consistency, to suggest new experiments and to interpret the results. They are powerful tools in basic research, integrating across several levels of analysis - from molecular, to synaptic, to cellular, to systems, to complex visual behavior.

[4]Note that some of the invariances may not depend on specific experience (e.g. learning how the appearance of a specific object varies) but on more general learning over basic visual features. For instance, the effects of 2D affine transformations, which consist of any combination of scaling, translation, shearing, and rotation in the image plane, can be estimated in principle from just one object view. Generic mechanisms in the system circuitry, independent of specific objects and object classes, can provide invariance to these transformations for all objects.

[5]The previous model implementation (see [Riesenhuber and Poggio, 1999b]) of the theory was sometimes referred to as *HMAX*. The theory described here is a significant extension of it – mainly because it includes learning stages in areas before IT – which was already planned in [Riesenhuber and Poggio, 1999b]. The early version of the model and the main differences with the present framework are listed in Appendix A.1.

[6]The theoretical work described in this paper started about 15 years ago with a simple model of view-based recognition of 3D objects [Poggio and Edelman, 1990], which in turn triggered psychophysical experiments [Bülthoff and Edelman, 1992] with paperclip objects previously introduced by Buelthoff and Edelman and then psychophysical [Logothetis et al., 1994] and physiological [Logothetis et al., 1995] exper-

iments in monkeys. The latter experiments triggered the development of a feedforward model [Riesenhuber and Poggio, 1999b] of the ventral stream to explain the selectivity and invariance found in IT neurons. The model, formerly known as HMAX, developed further into the theory described here as a direct effect of close interactions and collaborations (funded by NIH, DARPA and NSF) with several experimental groups, including David Ferster's, Earl Miller's, Jim DiCarlo's, Ilan Lampl's, Winrich Freiwald's, Marge Livingstone's, Ed Connor's, Aude Oliva's and David van Essen's, and with other close collaborators such as Max Riesenhuber, Christof Koch and Martin Giese.

[7]The hierarchical organization of visual cortex may be due to the need to build-in those invariances (to size, to position, to rotation) that do not need visual experience for the specific object but are valid for all objects (and could be – in part or completely – built into the DNA specifications for the architecture of visual cortex). This is not the case for illumination and viewpoint invariance that are the results of experience of images under different viewpoints and illuminations for *each specific* object. In fact, whereas view-invariant recognition requires visual experience of the specific novel object, position and scale invariance seems to be immediately present in the view-tuned neurons of IT cortex without the need of visual experience for views of the specific object at different positions and scales. It seems reasonable to assume that for object recognition hierarchies arise from a) the need to obtain selectivity and invariance within the constraints imposed by neuronal circuits, b) the need to learn from very few examples as biological organisms do and c) the need to reuse many of the same units for different recognition tasks.

2 Quantitative framework for the ventral stream

2.1 Feedforward architecture and operations in the ventral stream

The main physiological data summarized in the previous section, together with computational considerations on image invariances lead to a theory which summarizes and extends several previously existing models [Hubel and Wiesel, 1962, 1965b; Poggio and Edelman, 1990; Perrett and Oram, 1993; Mel, 1997; Wallis and Rolls, 1997; Riesenhuber and Poggio, 1999b, 2000; Elliffe et al., 2002; Thorpe, 2002; Amit and Mascaro, 2003] and biologically motivated computer vision approaches [Fukushima, 1980; Fukushima et al., 1983; Fukushima, 1986; LeCun, 1988; LeCun et al., 1998; Wersing and Koerner, 2003; LeCun et al., 2004]. The theory builds up on the classical Hubel & Wiesel model of simple and complex cells. We think that it represents the simplest class of models reflecting the known anatomical and biological constraints.

The theory maintains that:

1. One of the main functions of the ventral stream pathway is to achieve an exquisite trade-off between selectivity and invariance at the level of shape-tuned and invariant cells in IT from which many recognition tasks can be readily accomplished;

2. The underlying architecture is hierarchical, aiming in a series of stages to increasing invariance to object transformations and tuning to more specific *features*;

3. Two main functional types of units, *simple* and *complex*, represent the result of two main operations to achieve tuning (*S* layer) and invariance (*C* layer);

4. The two corresponding operations are a *normalized dot-product* – for (bell-shaped) Gaussian-like tuning of the simple units – and a *softmax* operation – for invariance (to some degree) to position, scale and clutter of the complex units.

The overall architecture is sketched in Fig. 2.1. We now describe its main features.

Mapping of computational architecture to visual areas. The model of Fig. 2.1 reflects the general organization of visual cortex in a series of layers from V1 to IT and PFC. The first stage of simple units (S1) – corresponding to the classical simple cells of Hubel & Wiesel – represents the result of a first tuning operation: Each S1 cell, receiving LGN (or equivalent) inputs, is tuned in a Gaussian-like way to a bar of a certain orientation among a few possible ones.

Each of the complex units (C1) in the second layer receives – within a neighborhood – the outputs of a group of simple units in the first layer at slightly different positions and sizes but with the same preferred orientation. The operation is a nonlinear softmax operation – where the activity of a pooling unit corresponds to the activity of the strongest input, pooled over a set of synaptic inputs. This increases invariance to local changes in position and scale while maintaining feature specificity.

At the next simple cell layer (S2), the units pool the activities of several complex (C1) with weights dictated by the unsupervised learning stage with different selectivities according to a Gaussian tuning function, thus yielding selectivity to more complex patterns with different selectivities by means of a (bell-shaped) Gaussian-like tuning function yielding selectivity to more complex patterns – such as combinations of oriented lines. The S2 receptive fields are obtained by this non-linear combination of C1 subunits.

Simple units in higher layers (S3 and S4) combine more and more complex features with a Gaussian tuning function, while the complex units (C2 and C3) pool their outputs through a max function providing increasing invariance to position and scale. In the model, the two layers alternate (though levels could be conceivably skipped, see [Riesenhuber and Poggio, 1999b]; it is likely that only units of the S type follow each other above C2 or C3). Also note that while the present implementation follows the hierarchy of the Fig. 2.1, the theory is fully compatible with a looser hierarchy.

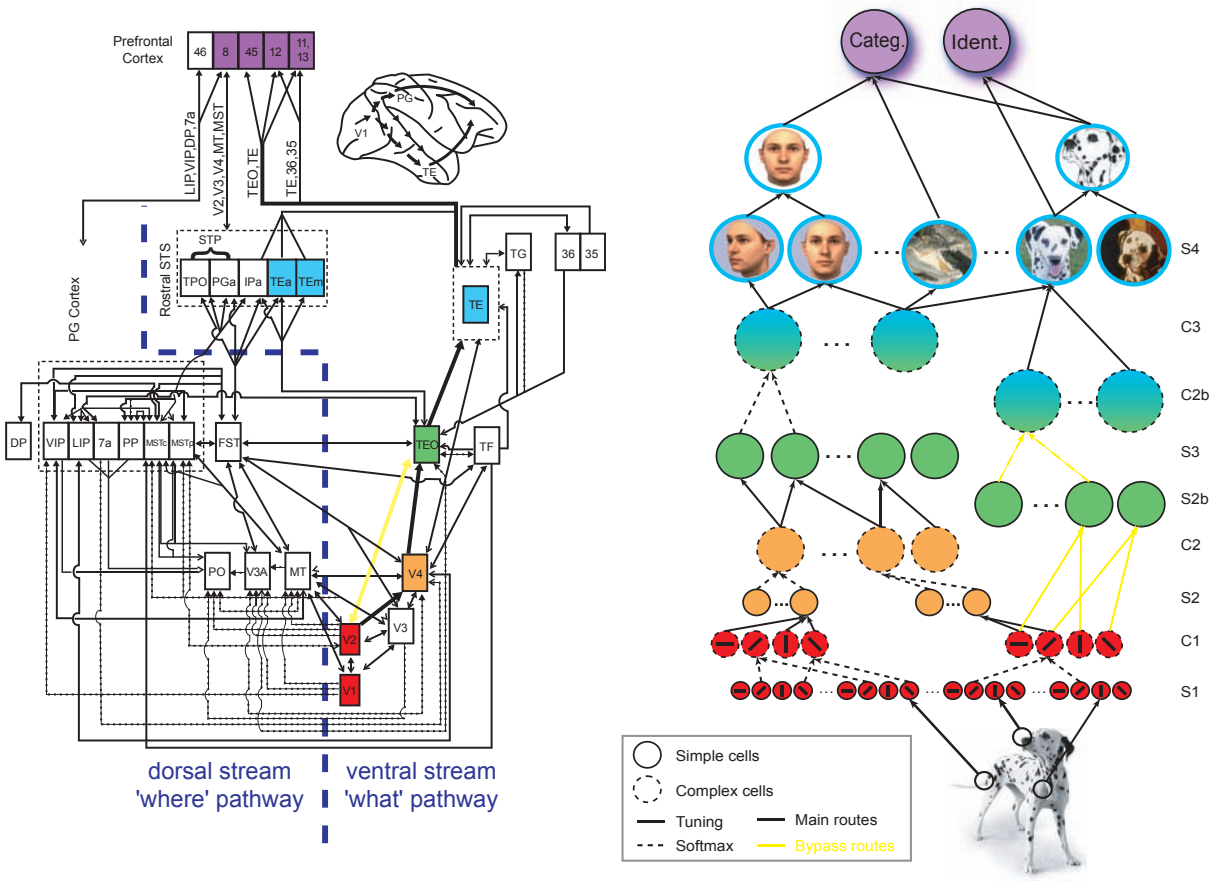


Figure 2.1: Tentative mapping between (right) functional primitives of the theory and (left) structural primitives of the ventral stream in the primate visual system (modifi edfrom Van Essen and Ungerleider [Gross, 1998]). Colors encode the correspondences between model layers and brain areas. Stages of *simple* units with Gaussian-like tuning (plain circles and arrows), which provide generalization [Poggio and Bizzi, 2004], are interleaved with layers of *complex* units (dotted circles and arrows), which perform a softmax operation on their inputs and provide invariance to position and scale (pooling over scales is not shown in the fi gure). Both operations may be performed by the same local recurrent circuits of lateral inhibition (see text). It is important to point out that the hierarchy is probably not as strict as depicted here. In addition there may be units with relatively complex receptive fi eldsalready in V1. The main route from the feedforward ventral pathway is denoted with black arrows while the bypass route [Nakamura et al., 1993] is denoted with yellow arrows. Learning in the *simple* unit layers from V4 up to IT (including the S4 view-tuned units) is assumed to be stimulus-driven (though not implemented at present, the same type of learning may be present in V1, determining receptive fi eldstuned to specifi csets of LGN-like subunits). It only depends on task-independent visual experience tuning of the units. Learning in the *complex* cell layers could, in principle, also be based on a task-independent trace rule exploiting temporal correlations (see [Földiák, 1991]). Supervised learning occurs at the level of the circuits in PFC (two sets of possible circuits for two of the many different recognition tasks – identifi cationand categorization – are indicated in the fi gure at the level of PFC). The model, which is feedforward (apart from local recurrent circuits), attempts to describe the initial stage of visual processing, *immediate recognition*, corresponding to the output of the top of the hierarchy and to the fi rst150 milliseconds in visual recognition.

In addition it is likely that the same stimulus-driven learning mechanism implemented for S2 and above (see later) operates also at the level of S1 units. In this case, we would expect S1 units with tuning not only for oriented bars but also for more complex patterns, corresponding to the combination of LGN-like, center-surround subunits. In any case, the theory predicts that the number of potentially active subunits (either of the LGN- or of the simple units- type later on) increases from V1 to IT. Correspondingly the size as well as the potential complexity of the receptive fields and the optimal tuning grow.

Two basic operations for selectivity and for invariance. The two key computational mechanisms in the model are: (a) tuning by the simple *S* units to build object-selectivity and (b) softmax by the complex *C* units to gain invariance to object transformations. The simple *S* units take their inputs from units that "look" at the same local neighborhood of the visual field but are tuned to *different preferred stimuli*. These subunits are combined with a normalized dot-product, yielding a bell-shaped tuning function, thus increasing object selectivity and tuning complexity.

The complex *C* units pool over inputs from *S* units tuned to the *same preferred stimuli* but at slightly different positions and scales through a softmax operation, thereby introducing tolerance to scale and translation. This gradual increase in both selectivity and scale is critical to avoid both a combinatorial explosion in the number of units, and the binding problem between features.

An approximative – and static – mathematical description of the two operations is given below – though the most precise definition will be in terms of underlying biophysical micro-circuits (see Section 5, which also provides the dynamics of the processing). The tuning operation, represented in simple *S* units, is provided by:

$$y = g\left(\frac{\sum_{j=1}^{n} w_j \, x_j^p}{k + \left(\sum_{j=1}^{n} x_j^q\right)^r}\right), \tag{1}$$

where x_i is the response of the i-th pre-synaptic unit, w_i the synaptic strength of the connection between the i^{th} pre-synaptic unit and the simple unit y, k a constant (set to a small value to avoid zero-divisions) and g is a sigmoid transfer function that controls the sharpness of tuning such that:

$$g(t) = 1/(1 + e^{\alpha(t-\beta)}). \tag{2}$$

The exponents p, q and r represent the static nonlinearities in the underlying neural circuit. Such nonlinearity may arise from the sigmoid-like threshold transfer function of neurons, and depending on its operating range, various degrees of nonlinearities can be obtained. Fig. 2.2 indicates, along with a plausible circuit for the tuning operation, possible locations for the synaptic nonlinearities p, q, and r, from Eq. 1.

In general, when $p \leq qr$, y has a peak around some value proportional to the w_i's, that is (as in the classical perceptron) when the input vector $\mathbf{x}$ is collinear to the weight vector $\mathbf{w}$. For instance, when $p = qr$ (*e.g.*, $p = 1$, $q = 2$ and $r = 1/2$), a tuning behavior can be obtained by adding a fixed bias term b (see Appendix A.4) and [Maruyama et al., 1991, 1992; Kouh and Poggio, 2004]). Even with $r = 1$ and $p \approx q$, tuning behavior can be obtained if $p < q$ (*e.g.*, $p = 1$ and $q = 1.2$ or $p = 0.8$ and $q = 1$). Note that the tuning is determined by the synaptic weights $\mathbf{w}$. The vector $\mathbf{w}$ determines – and corresponds to – the *optimal* or *preferred stimulus* for the cell.

The softmax operation, represented in *complex* units, is given by:

$$y = g\left(\frac{\sum_{j=1}^{n} x_j^{q+1}}{k + \left(\sum_{j=1}^{n} x_j^q\right)}\right), \tag{3}$$

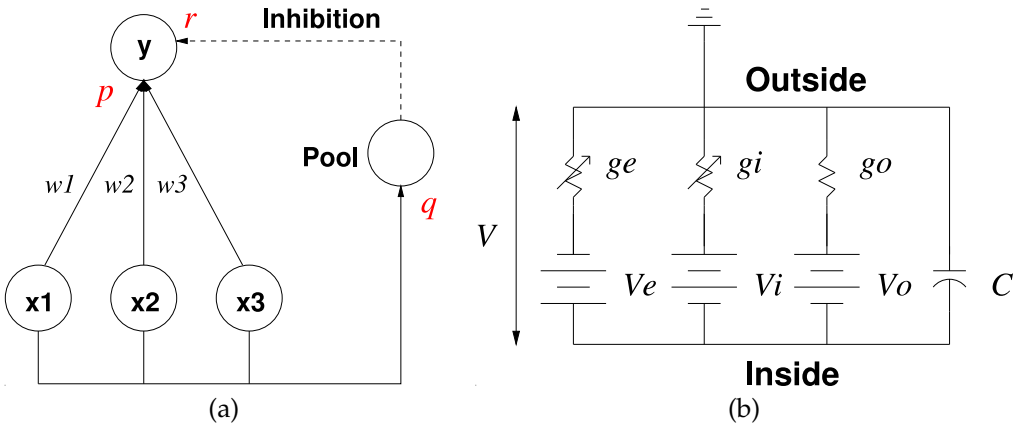


Figure 2.2: (a) One possible neural circuitry that performs divisive normalization and weighted sum. Depending on different degrees of nonlinearity in the circuit (denoted by p, q and r), the output y may be a tuning (Eq. 1) or an invariance (Eq. 3) operation. The gain control, or the normalization mechanism in this case is achieved by a feedforward shunting inhibition. Such circuit has been proposed in [Torre and Poggio, 1978; Reichardt et al., 1983; Carandini and Heeger, 1994]. The effects of shunting inhibition can be calculated from a circuit diagram like (b), see Section 5.

which is of the same form as the tuning operation (assume $w_i = 1$ and $p = q + 1$ in Eq. 1). For sufficiently large q, this function approximates a maximum operation (softmax, see [Yu et al., 2002]). In particular, Eq. 1 yields tuning for $r = 1$, $p = 1$ $q = 2$ and an approximation of max for $r = 1$, $p = 2$ $q = 1$.

As we mentioned earlier, the description above is a static approximation of the dynamical operation implemented by the plausible neural circuits described in Section 5; it is however the operation used in most of the simulations described in this paper. One could give an even simpler – and more idealized – description, which was used in earlier simulations [Riesenhuber and Poggio, 1999b] and which is essentially equivalent in terms of the overall properties of the model.

In this description the tuning operation is approximated in terms of a multivariate Gaussian (which can indeed approximate a normalized dot-product well – in a higher dimensional space, see [Maruyama et al., 1991]):

$$y = \exp^{-\frac{\sum_{j=1}^{n}(x_j - w_j)^2}{2\sigma^2}} \tag{4}$$

where σ characterizes the width (or tuning bandwidth) of the Gaussian centered on $\mathbf{w}$.

Similarly, in this simplified description, the softmax operation is described as a pure max that is:

$$y = \max_{i \in \mathcal{N}} x_i. \tag{5}$$

Despite the fact that a $\max$ operation seems very different from Gaussian tuning, Eq. 1 and Eq. 3 are remarkably similar. Their similarity suggests similar circuits for implementing them, as discussed in Section 5 and shown in Fig. 2.2.

Notes

[1] In Eq. 1 the denominator in principle involves "all" inputs in the neighborhood, even the ones with synaptic weights set to zero. For example, for simple units in V1, the denominator will include "all" LGN inputs, not only the ones actually contributing to the excitatory component of the activity of the specific simple unit. As a consequence, the denominator could normalize the activity across simple units in V1. Studies such as [Cavanaugh et al., 2002] suggest that the normalization pool is probably larger than the classical receptive field. In our present implementation both the numerator and denominator were limited to the inputs within the "classical" receptive field of the specific simple unit, but we plan to change this in future implementations using tentative estimates from physiology.

[2]Note that a more general form of normalization in Eq. 1 would involve another set of synaptic weights $\tilde{\mathbf{w}}$ in the denominator, as explored in a few different contexts such as to increase the independence of correlated signals [Heeger et al., 1996; Schwartz and Simoncelli, 2001] and the biased competition model of attention [Reynolds et al., 1999].

[3]In Eq. 3, when $p > qr$ (this condition is the opposite of the tuning function) and the output is scaled by a sigmoid transfer function, the softmax operation behaves like a winner-take-all operation.

[4]A prediction of the theory is that tuning and normalization are tightly interleaved. Tuning may be the main goal of normalization and gain control mechanisms throughout visual cortex.

[5]An alternative to the tuning operation based on the *sigmoid of a normalized dot-product* (see Eq. 1) is a *sigmoid of a dot-product* that is

$$y = g\left(\sum_{j=1}^{n} w_j \, x_j^p\right), \tag{6}$$

where g is the sigmoid function given in Eq. 2. Eq. 6 is less flexible than Eq. 1, which may be tuned to any arbitrary pattern of activations, even when the magnitudes of those activations are small. On the other hand, a neural circuit for the dot-product tuning does not require the inhibitory elements and, thus, is simpler to build. Also, with a very large number of inputs (high dimensional tuning), the total activation of the normalization pool, or the denominator in Eq. 1, would be more or less constant for different input patterns, and hence, the dot product tuning would work like the normalized dot product. In other words, the normalization operation would only be necessary to build a robust tuning behavior with a small number of inputs. It is conceivable that both Eq. 1 and Eq. 6 are used for tuning, with Eq. 6 more likely in later stages of the visual pathway.

[6]A comparison of Gaussian tuning, normalized dot-products and dot-product with respect to the recognition performance of the model is described in Appendix A.4 and A.5. A short summary is that the three similar tuning operations yield similar model performance on most tests we have performed so far.

[7]The model is quite robust with respect to the precision of the tuning and the max operations. This is described in the Appendix A.5, which examines, in particular, an extreme quantization on the activity of the S4 inputs (*i.e.*, binarization, which is the case of single spiking axons – the "thin" cables of Section 5 – over a short time intervals).

2.2 Learning

Various lines of evidence suggest that visual experience – during and after development – together with genetic factors determine the connectivity and functional properties of units. In the theory we assume that learning plays a key role in determining the wiring and the synaptic weights for the S and the C layers. More specifically, we assume that the tuning properties of simple units – at various levels in the hierarchy – correspond to learning combinations of "features" that appear most frequently in images. This is roughly equivalent to learning a dictionary of patterns that appear with high probability. The wiring of complex units on the other hand would reflect learning from visual experience to associate frequent transformations in time – such as translation and scale – of specific complex features coded by simple units. Thus learning at the S and C level is effectively *learning correlations* present in the visual world.

The S layers' wiring depends on learning correlations of features in the image at the *same time*; the C layers' wiring reflects learning correlations *across time*. Thus the tuning of simple units arises from learning correlations in space (for S1 units the bar-like arrangements of LGN inputs, for S2 units more complex arrangements of bar-like subunits, *etc.*). The connectivity of complex units arises from learning correlations over time, *e.g.*, that simple units with the same orientation and neighboring locations should be wired together in a complex unit because often such a pattern changes smoothly in time (*e.g.*, under translation).

Figure 2.3: Sample natural images used to passively expose the model and tune units in intermediate layers to the statistics of natural images.

At present it is not clear whether these two types of learning would require two different types of synaptic "rules" or whether the same synaptic mechanisms for plasticity may be responsible through slightly different circuits, one involving an effective time delay. A third type of synaptic rule would be required for the task-dependent, supervised learning at the top level of the model (tentatively identified with PFC).

In this paper we have quantitatively simulated learning at the level of simple units only (at the S2 level and higher). We plan to simulate learning of the wiring at the level of C units in later work using the "trace" rule (see [Földiák, 1991; Rolls and Deco, 2002] for a plausibility proof). We also plan to simulate learning – experimenting with a few plausible learning rules – from natural images of the tuning of S1 units, where we expect to find mostly directional tuning *but also more complex types of tuning*. We should emphasize that the implementation of learning in the model is at this point more open-ended than other parts of it because little is known about learning mechanisms in visual cortex.

2.2.1 Learning a universal dictionary of shape-tuned (S) units: from S2 to S4 (V4 to AIT)

The tuning properties of neurons in the ventral stream of visual cortex, from V1 to inferotemporal cortex (IT), play a key role for visual perception in primates and in particular for their object recognition abilities. As we mentioned, the tuning of specific neurons probably depends, at least in part, on visual experience. In the original implementation of the model [Riesenhuber and Poggio, 1999b], learning only occurred in the top-most layers of the model (*i.e.*, units corresponding to the view-tuned units in AIT [Logothetis et al., 1995] and the task-specific circuits from IT to PFC [Freedman et al., 2001]). Because the model was initially tested on simplified stimuli (such as paperclips or faces on a uniform background), it was possible to manually tune units in intermediate layers (simple 2×2 combinations of 4 orientations) [Riesenhuber and Poggio, 1999b] to be selective for the target object.

The original model did not perform well on large datasets of real-world images (such as faces with different illuminations, background, expression, *etc.*) [Serre et al., 2002; Louie, 2003]. Consistent with the goals of the original theory [Riesenhuber and Poggio, 1999b], we describe here a simple biologically plausible learning rule that determines the tuning of S units from passive visual experience. The learning rule effectively generates a *universal and redundant dictionary* of shape-tuned units from V4 to IT that could, in principle, handle several visual recognition tasks (*e.g.*, face identification ("who is it?"), face categorization (''is it a face?"), as well as gender and expression recognition, *etc.*).

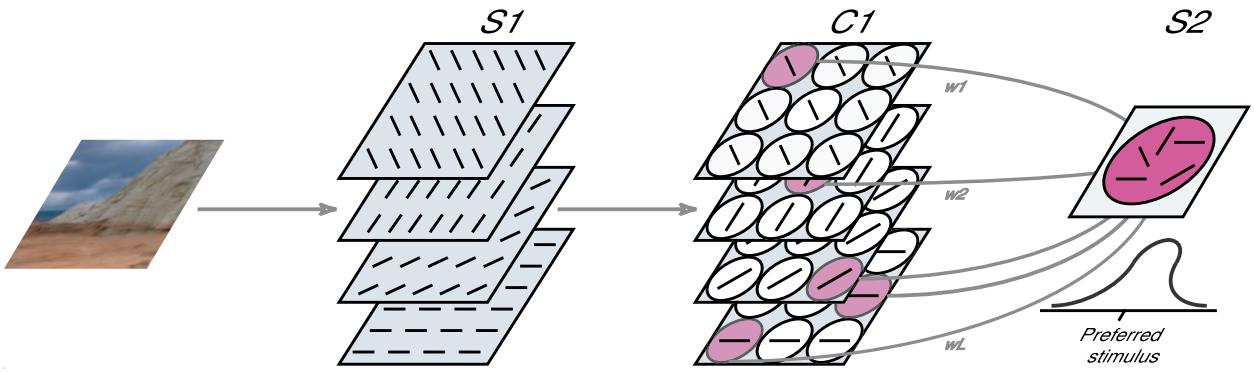


Figure 2.4: Imprinting process at the S_2 level to generate a universal dictionary of shape-tuned units.

The rule we assume is very simple though we believe it will have to be modified somewhat to be biologically more plausible. Here we assume that during development neurons in intermediate brain areas become tuned to the pattern of neural activity induced in the previous layer by natural images. This is a reasonable assumption considering recent theoretical work that has shown that neurons in primary visual cortex are tuned to the statistics of natural images [Olshausen and Field, 1996; Hyvärinen and Hoyer, 2001].

During training each unit in the simple layers (S2, S2b and S3 sequentially) becomes tuned by exposing the model to a set of 1,000 natural images unrelated to any categorization task. This image dataset was collected from the web and includes different natural and artificial scenes, *e.g.,* landscapes, streets, animals, *etc.*; see Fig. 2.3). For each image presentation, starting with the S2 layer, some units become tuned to the pattern of activity of their afferents. This training process can be regarded as an imprinting process in which each *S* unit (*e.g.,* S2 unit) stores in its synaptic weights the specific pattern of activity from its afferents (*e.g.,* C1 units) in response to the part of the natural image that falls within its receptive field. A biologically plausible version of this rule could involve mechanisms such as LTP. In the model this is done by setting, for each unit-type, the vector $\mathbf{w}$ in Eq. 1 (or equivalently Eq. 4) to the pattern of pre-synaptic activity. For instance, when training the S2 layer, this corresponds to setting $\mathbf{w}$ to be equal to the activity of the C1 afferent units $\mathbf{x}$. The unit becomes *tuned* to the particular stimulus presented during learning, *i.e.*, the part of the natural image that fell within its receptive field during learning, which in turn becomes the preferred stimulus of the unit. That is, the unit response is now maximal when a new input $\mathbf{x}$ matches exactly the learned pattern $\mathbf{w}$ and decreases (with a bell-shape profile) as the new input becomes more dissimilar. Fig. 2.4 sketches this *imprinting* process for the S2 layer. Learning at the S2b and S3 level takes place in a similar way.

We assumed that the images move (shifting and looming) so that each type of *S* unit is being replicated across the visual field. The tuning of units from S1 to C3 is fixed after this development-like stage. Afterward, only the task-specific circuits from IT to PFC required learning for the recognition of specific objects and object categories. It is important to point out that this *same universal* dictionary of shape-tuned units (up to S4) is later used to perform the recognition of many different object categories (*e.g.*, animals, cars, faces,*etc.* (see Section 3).

2.2.2 Task-dependent learning: from IT to PFC

As discussed in the introduction (see also the discussion), we assume that a particular "program" is set up – probably in PFC – depending on the task. In a passive state (no specific visual task is set) there may be a default routine running (perhaps the routine: *what is there?*). For this paper, we think of the routine running in PFC as a (linear) classifier trained on a particular task in a supervised way and "looking" at the activity of a few hundred neurons in IT. Note that a network comprising units in IT with a Gaussian-like tuning function together with a linear classifier on their outputs, is equivalent to a regularized RBF classifier, which is among the most powerful in terms of learning to generalize [Poggio and Bizzi, 2004].

Interestingly, independent work [Hung et al., 2005a] demonstrated that linear classifiers can indeed *read-out* with high accuracy and over extremely short times (a single bin as short as 12.5 millisecond) object identity, object category and other information (such as position and size of the object) from the activity of about 100 neurons in IT (see Section 4.3.3).

When training the model to perform a particular recognition task, like the animal *vs.* non-animal categorization task presented in Section 3.2 for instance, S4 units (corresponding to the view-tuned units in IT) are imprinted with examples from the training set (*e.g.*, animal and non-animal stimuli in the animal detection task, car and non-car stimuli for car detection, *etc.*). As for the S2, S2b and S3 units, imprinting of a particular S4 unit consists in storing in its synaptic weights $\mathbf{w}$ the precise pattern of activity from its afferents (*i.e.*, C3 and C2b units with different invariance properties) by a particular stimulus. Again the reader can refer to Fig. 2.4 for an example of imprinting. In all simulations presented in this paper, we only used $\frac{1}{4}$ of the entire stimulus set available for training to imprint the S4 units, while we used the full set to learn the synaptic weights of the linear classifier that "looks" at IT.

The linear classifier from IT to PFC used in the simulations corresponds to a supervised learning stage with the form:

$$f(\mathbf{x}) = \sum_i c_i K(\mathbf{x^i}, \mathbf{x}) \quad \text{where} \quad K(\mathbf{x^i}, \mathbf{x}) = g\left(\frac{\sum_{j=1}^{n} w_j \ (x_j^i)^p}{k + \left(\sum_{j=1}^{n} (x_j^i)^q \right)^r} \right) \tag{7}$$

characterizes the response of the i^{th} S4 unit tuned to the training example $\mathbf{x^i}$ (animal or non-animal) to the input image $\mathbf{x}$ and $\mathbf{c}$ is the vector of synaptic weights from IT to PFC. The superscript i indicates the index of the image in the training set and the subscript j indicates the index of the pre-synaptic unit. Since the S4 units (corresponding to the view-tuned units in IT) are like Gaussian radial basis functions (RBFs), the part of the network in Fig. 2.1 comprising the inputs to the S4 units up to PFC can be regarded as an RBF network (see Appendix A.6)). Supervised learning at this stage involves adjusting the synaptic weights $\mathbf{c}$ so as to minimize a (regularized) error on the training set. The reader can refer to Appendix A.6 for complementary remarks and connections between RBF networks and cortical circuits in PFC.

2.2.3 Training the model to become an expert by selecting features for a specific set of objects

To improve further performance on a *specific task* it is possible to select a subset of the S units that are selective for a target-object class (*e.g.*, face) among the *universal* dictionary – *e.g.*, the very large set of all S units learned from natural images. This step can be applied on all S2, S2b and S3 units (sequentially) so that, at the top of the hierarchy, the view-tuned units (S4) receive inputs from object-selective afferents only.

This step is still task independent as the same basic units (S2, S2b, S3 and S4 units) would be used for different tasks (*e.g.*, face- detection, -identification, or gender-recognition). This learning step improves performance with respect to using the "universal" dictionary of features extracted from natural images (see Fig. 2.3) but the latter already produce excellent performances in a variety of tasks. Thus selection of specific features is not strictly needed and certainly is not necessary initially for recognition tasks with novel sets of objects.

Our preliminary computational experiments suggest that performance on specific recognition tasks when only a few (supervised, *e.g.*, an input pattern $\mathbf{x}$ and its label $\mathbf{y}$ pairs) examples of the new task are available is higher if the *universal dictionary is used*; on the other hand, when the number of labeled examples increase, selection of tuned units appropriate for the task *increases* performance. In summary, an organism would do fine in using the universal features in the initial phase of dealing with a new task but would do better by later selecting appropriate features (*e.g.*, tuned units) out of this dictionary when more experience with the task becomes available.

The proposed approach seek *good* units to represent the target-object class, *i.e.*, units that are robust to target-object transformations (*e.g.*, inter-individuals variations, lighting, pose, *etc.*). A detailed presentation of the learning algorithm is provided in Appendix A.3. By presenting the model with sequences of images that contain the target-object embedded in various backgrounds (see Fig. 3.2 for typical stimuli used), the algorithm selects features that are robust against clutter and within-class shape variations. In another setting, the model could be exposed to an image sequence of the target-object (*e.g.*, a face) undergoing a series of transformations (*e.g.*, a face rotating in depth to learn a pose-invariant representation). Note that learning here is unsupervised and the only constraint is for the model to be presented with a short image sequence containing the target-object in isolation. [8]

Notes

[8]While we did not test for tolerance to distractors, we expect the learning scheme to be robust to the short presentation of distractors.

[9]It should be clear that the model is a convenient way to summarize known facts and to ask through quantitative computer simulations a number of questions relevant for experiments; it is very much a working hypothesis – a framework and a computational tool – rather than a complete, finished model. It will certainly change – possibly in a drastic way – as an effect of the interaction with the psychophysical and physiological experiments.

[10]As described in Table 1, the model contains on the order of 10^7 units (these bounds are computed using reasonable estimates for the S4 receptive field sizes and the number of different types of simple units in all S layers). This number may need to be increased by no more than one or two orders or magnitude to obtain an estimate of the number of biological neurons which are needed – based on the circuits described in Section 5 and speculations on the "cable hypothesis" (see Section 5). This estimate results in about $10^8 - 10^9$ actual neurons, which corresponds to about 0.01% to 1% of visual cortex (based on 10^{11} neurons in cortex [Kandel et al., 2000]). This number is far smaller than the proportion of cortex taken by visual areas.

[11]We shall emphasize that, even though the number above was computed for a version of the model trained to perform a single (binary) animal *vs.* non-animal classification task – because the same basic dictionary of shape-tuned units (*i.e.,* from S1 up to S4) is being used for different recognition tasks – this number would not change significantly for a more realistic number of categories. In particular, training the model to recognize a plausible number of discriminable objects (*i.e.,* probably no more than $30,000$ [Biederman, 1987]), would add only an extra 10^7 S4 units.

[12]It should be emphasized that the various layers in the architecture – from V1 to IT – create a large redundant dictionary of features with different degrees of selectivity and invariance. It may be advantageous for circuits in later areas (say classifier circuits in PFC) to have access not only to the highly invariant and selective units of AIT but also to less invariant and simpler units of the V2 and V4 type. For instance recent work (in submission) by Bileschi & Wolf has shown that the performance of the model on the 101 object database (see section 3 and [Serre et al., 2005c]) containing different objects with large variations in shape but limited ranges of positions and scales could be further improved by 1) restricting the range of invariance of the top units and 2) passing some of the C1 unit responses to the classifier along with the top unit responses. We also found in the animal *vs.* non-animal categorization task of Section 3 that performance were improved with S4 units that not only received their inputs from the top (C3) units but also from low-level C1 units (with limited invariance to position and scale) and C2b units (of intermediate complexity with some range of invariance). Finally preliminary computational experiments by Meyers & Wolf suggest for instance that "fine" recognition tasks (such as face identification) may benefit from using C1 inputs *vs.* S4 inputs. In cortex there exist at least two ways by which the response from lower stages could be incorporated in the classification process: 1) Through bypass routes [Nakamura et al., 1993] (for instance through direct projections between intermediate areas and PFC) and / or 2) by replicating some of the unit types from one layer to the next. This would suggest the existence of cells such as V1 complex cells along with the bulk of cells in the various stages of visual cortex. We are of course aware of the potential implications of observation (1) for how back-projections could gate and control inputs from lower areas to PFC in order to optimize performance in a specific task. From the same point of view, direct connections from lower visual areas to PFC make sense.

Layers	Number of units
S1	1.6×10^6
C1	2.0×10^4
S2	1.0×10^7
C2	2.8×10^5
S3	7.4×10^4
C3	1.0×10^4
S4	1.5×10^2
S2b	1.0×10^7
C2b	2.0×10^3
Total	2.3×10^7

Table 1: Number of units in the model. The number of units in each layers was calculated for the animal *vs.* non-animal categorization task presented in Section 3.2, *i.e.,* S4 (IT) receptive fields (RF) of 4.4^o of visual angle (160×160 pixels, probably not quite matching the number of photoreceptors in the macaque monkey in that foveal area of the retina) and about $2,000$ unit types in each S2, S2b and S3 layers.

3 Performance on Natural Images

In this section we report results obtained with the model on natural image databases. We first show that the model can handle the recognition of many different object categories presenting results on a database of 101 different objects [Fei-Fei et al., 2004] from the Caltech vision group. We also briefly summarize results that show than the model outperforms several other benchmark AI systems (see [Serre et al., 2004, 2005c] for details). Most importantly we show that the model perform at human level on a rapid animal / non-animal recognition task.

3.1 Comparison with state-of-the-art AI systems on different object categories

To evaluate the plausibility of the theory, we trained and tested the model on a real-world object recognition database, called the *Caltech 101-object category database* [Fei-Fei et al., 2004]. The database contains 101 different object categories as well as a "background" category that does not contain any object. Images were collected from the web using various search engines. This is a challenging database as it contains images from many different object-categories with variations in shapes, clutter, pose, *etc.*Importantly the set of images used from training the model was unsegmented, *i.e.,* the target-object was embedded in clutter. The database is publicly available at `http://vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html`.

The model was trained as indicated in Subsection 2.2. First, for each object category, S units (S2, S2b and S3) that are selective for the target-object class were selected using the *TR* learning algorithm (see Appendix A.3 from image sequences artificially generated with examples selected from a training set of images. In a second step S4 (view-tuned) units, receiving inputs from C2b and C3 units, were imprinted with examples from the training set (that contains stimuli from both the target object class as well as distractors from the "background" set) and a linear classifier corresponding to task-specific circuits running between AIT and PFC was trained to perform an object present / absent recognition task (*e.g.,* face *vs.* non-face, rooster *vs.* non-rooster, *etc.*).

Fig. 3.1 shows typical results obtained with the model on various object categories. With only a few training examples (40 positive and 50 negative), the model is able to achieve high recognition rates on many different object categories.

We also performed a systematic comparison between the model and other benchmark AI systems from our own group as well as the Caltech vision group. For this comparison, we used five standard (Caltech) datasets publicly available (airplanes, faces, leaves, cars and motorcycles) and two MIT-CBCL datasets (faces and cars). Details about the datasets and the experimental procedure can be found in [Serre et al., 2005c, 2004]. For this comparison we only considered a sub-component of the model corresponding to the *bypass route*, *i.e.,* the route projecting directly from V1/V2 (S1/C1) to IT (S2b/C2b) thus bypassing V4, see Fig. 2.1. This was shown to constitute a good compromise between speed and accuracy for this application oriented toward AI and computer vision. We also replaced the final classification stage (equivalent to an RBF scheme [Poggio and Bizzi, 2004] in the model, which includes the S4 units in IT and the task-specific circuits from IT to PFC) with a more standard "computer vision classifier" (Ada boost). This allowed for a more rigorous comparison at the representation-level (model C2b units *vs.* computer vision features such as SIFT [Lowe, 1999], component-experts [Heisele et al., 2001; Fergus et al., 2003; Fei-Fei et al., 2004], or fragments [Ullman et al., 2002; Torralba et al., 2004]) rather than at the classifier level.

Fig. 3.2 shows typical examples from the five Caltech datasets for comparison with the constellation models [Weber et al., 2000; Fergus et al., 2003]. Table 2 summaries our main results. The model performs surprisingly well, better than all the systems we have so far considered. This level of performance was observed for object recognition in clutter, for segmented object recognition and for texture recognition. Extensive comparison with several state-of-the-art computer vision systems on several image dataset can be found in [Serre et al., 2004, 2005c; Bileschi and Wolf, 2005]. The source code that we used to run those simulation is freely available online at `http://cbcl.mit.edu/software-datasets`.

Notes

[1]We would like to thank Max Riesenhuber, Jennifer Louie, Rodrigo Sigala and Robert Liu for their contribution in earlier phases of this work.

[2]The model evaluation on real-world image datasets was done in collaborations with Lior Wolf and Stan Bileschi.

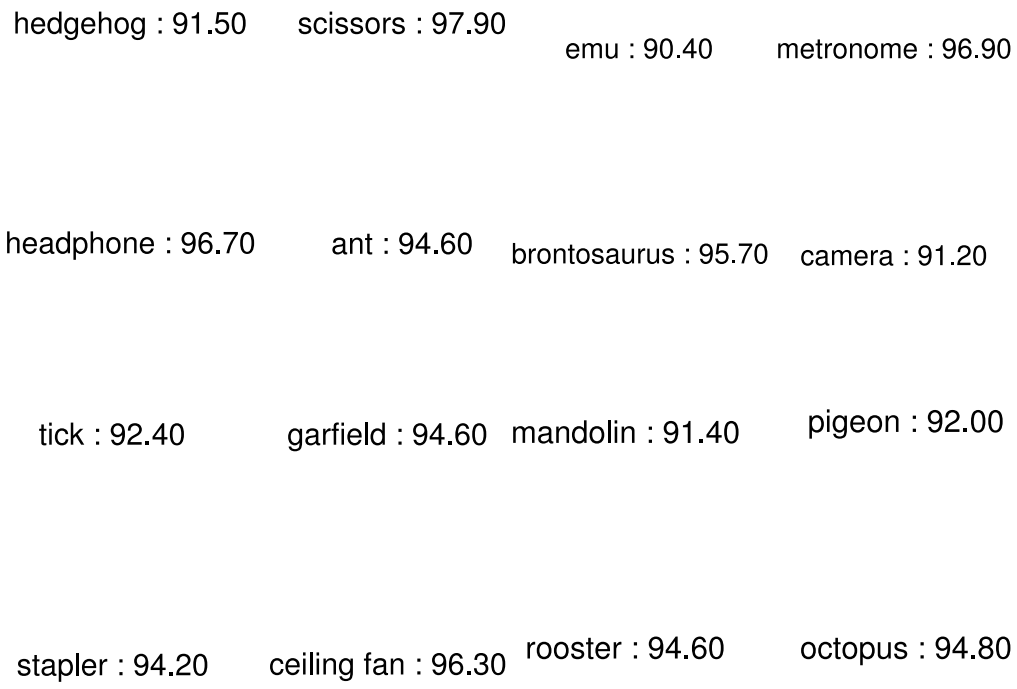


Figure 3.1: The model can handle many different recognition tasks. Here we show typical results from the 101-object database [Fei-Fei et al., 2004]. The model performance is indicated above each thumbnail, which represents a typical example from the database. The performance was evaluated on a binary classification task, *i.e.,* object present or absent (the set of distractors was chosen from the database "background" category as in [Fei-Fei et al., 2004]). Each number is the average of 10 random runs were the model was trained with only 50 negative and 40 positive training examples selected at random and tested on 50 positive and 50 negative examples. The error measure is the roc area (*i.e.,* the area under the roc curve). The experimental procedure was the same as in [Fei-Fei et al., 2004; Serre et al., 2005c].

Datasets			AI systems	Model
(CalTech)	Leaves	[Weber et al., 2000]	84.0	97.0
(CalTech)	Cars	[Fergus et al., 2003]	84.8	99.7
(CalTech)	Faces	[Fergus et al., 2003]	96.4	98.2
(CalTech)	Airplanes	[Fergus et al., 2003]	94.0	96.7
(CalTech)	Motorcycles	[Fergus et al., 2003]	95.0	98.0
(MIT-CBCL)	Faces	[Heisele et al., 2001]	90.4	95.9
(MIT-CBCL)	Cars	[Leung, 2004]	75.4	95.1

Table 2: Model *vs.* benchmark AI recognition systems. For the five Caltech datasets (leaves, cars, faces, airplanes, motorbikes), where the task is object recognition in clutter we compare against benchmarks which are based on the part-based generative model termed the constellation model [Weber et al., 2000; Fergus et al., 2003]. For the MIT-CBCL face dataset we compare with a hierarchical SVM-based architecture that was by itself shown to outperform many other face-detection systems [Heisele et al., 2001]. For the MIT-CBCL car dataset we compared to a system by [Leung, 2004] that uses fragments [Ullman et al., 2002] and AdaBoost.

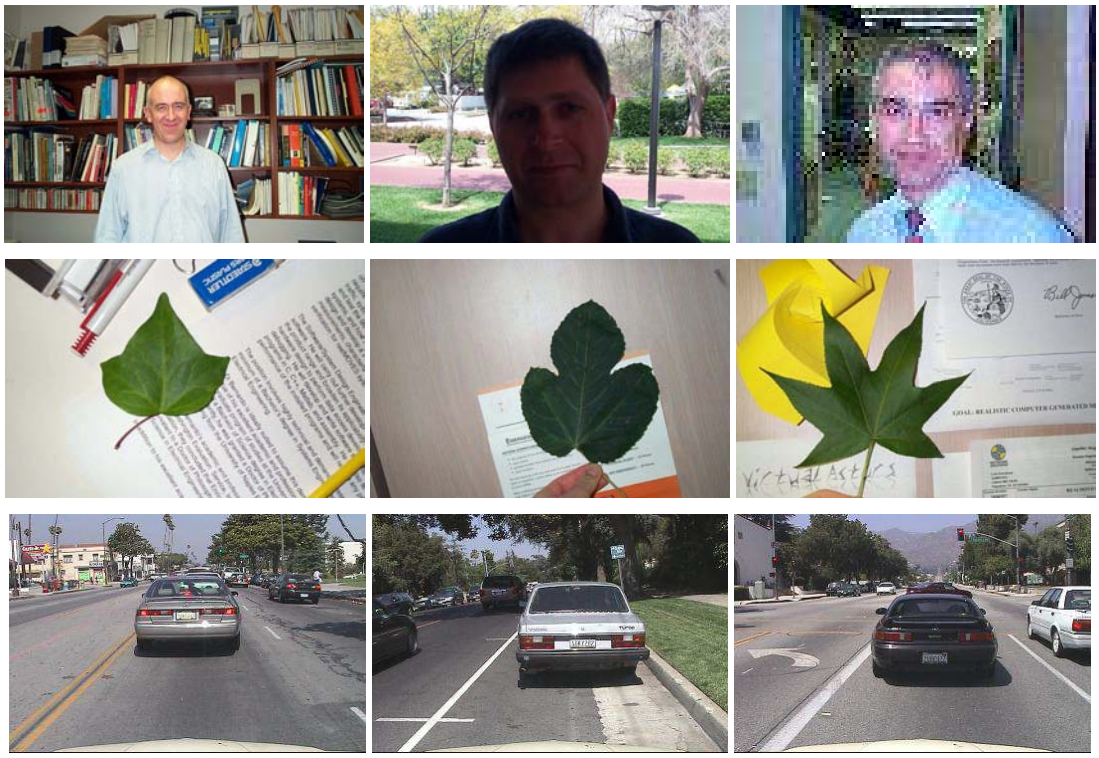


Figure 3.2: Sample images from the five Caltech datasets. From top to bottom: airplanes, motorcycles, faces, leaves and cars. Note that no color information is being used by the present version of the model to perform the task.

3.2 Predicting human performance on a rapid-categorization task

As we showed in the previous section, the theory provides a model that is capable of recognizing well complex images, *i.e.*, when tested on real-world natural images, a quantitative model implementing the theory, competes with and may even outperform state-of-the-art computer vision systems on several categorization tasks (see also [Serre et al., 2005c,b]). This is quite surprising, given the many specific biological constraints that the theory had to satisfy. It remained however still unclear whether any feedforward model could duplicate human performance in natural recognition tasks. Below we show (*in collaboration with Aude Oliva* [Serre et al., 2005a]) that indeed the model performs at human level on a rapid animal / non-animal categorization task [Thorpe et al., 1996], a task for which it is believed that feedback loops do not play a major role.

3.3 Immediate recognition and feedforward architecture

All feedforward proponents recognize that normal, everyday vision includes top-down effects mediated by the extensive anatomical back-projections found throughout visual cortex. Back-projections are likely to effectively run "programs" for reading out specific task-dependent information from IT (*e.g.,* is the object in the scene an animal? Or what is the approximate size of the object in the image [Hung et al., 2005a]?) They could also run specific routines in areas lower than IT - possibly by tuning or changing synaptic connection weights. During normal vision, back-projections are likely to dynamically control routines running at all levels of the visual system - throughout fixations and attentional shifts. Thus the key claim of feedforward models is that the first 100-200 milliseconds of visual perception involves mainly feedforward processing and that, during that time, complex recognition tasks are mostly accomplished by feedforward computations. In the same way as an experimental test of Newton's second law requires choosing a situation in which friction is negligible, we looked for an experimental paradigm in which recognition has to be fast and cortical back-projections are likely to be inactive. The paradigm we chose is ultra-rapid object categorization.

Ultra-rapid object categorization [Thorpe et al., 1996] likely depends only on feedforward processing [Thorpe et al., 1996; Thorpe and Fabre-Thorpe, 2001; van Rullen and Koch, 2003b]. Human observers can discriminate a scene that contains a particular prominent object, like an animal or a vehicle after only 20 milliseconds of exposure; ERP components related to either low-level image features of the image categories (*e.g.,* animal or vehicles) or to the image status (animal present or absent) are available at 80 and 150 milliseconds respectively. These experimental results establish an upper bound on how fast categorical decisions can be made by the human visual system, and suggest that categorical decisions can be implemented within a feed-forward mechanism of information processing [Thorpe et al., 1996; Thorpe and Fabre-Thorpe, 2001; van Rullen and Koch, 2003b; Keysers et al., 2001].

3.4 Theory and humans

We selected [Serre et al., 2005a] a set of balanced animal and non-animal stimuli from a commercially available database (Corel Photodisc) as in [Thorpe et al., 1996]. Animal stimuli are a rich class of stimuli as they offer a large variety in texture, shape, size, etc. We selected and grouped six hundred animal images into four categories, each category corresponding to a different viewing-distance from the camera: *heads* (close-ups), *close-body* (animal body occupying the whole image), *medium-body* (animal in scene context) and *far-body* (small animal or groups of animals in larger context). To make the task harder and prevent subjects from relying on low-level cues such as image-depth, we carefully selected a set of six hundred distractor images to match each of the four viewing-distances. Distractor images were of two types (three hundred of each): artificial or natural scenes. Images were all converted to gray values. This is because a) the model does not use color information, b) it has been previously shown that color is not diagnostic for rapid animal vs. non-animal categorization task [Delorme et al., 2000] and c) this is easier to mask. Fig. 3.4 (inset) shows typical examples of the stimuli used in this experiment.

The model was trained in two steps as described in Section 2. First, in a task-independent stage, a redundant dictionary of shape-tuned units (from V4 to IT, see Fig. 2.1) is learned by passively exposing the model with thousands of patches from natural images. During this stage S2 and S3 units are imprinted with patches of natural images, which become their preferred stimuli. The second learning stage consists in training the task-specific circuits, which may correspond to a routine running in PFC as a (linear) classifier trained on a particular task in a supervised way and looking at the activity of a few hundred neurons in IT. The final classifier was trained using 50 random splits on the stimuli database. In a typical run, half of the images were selected at random for training and the model was tested on the remaining ones. Leave-out evaluations of this kind have been shown to leave the best unbiased performance estimates. The model performance was surprisingly good given the difficulty of the task (large variations in shape, size, position, clutter) and relatively small training set.

In short images were briefly flashed for 20 ms, followed by an inter-stimulus interval of 30 ms, followed by a mask (80 ms, $1/f$ noise). The parameters were chosen according to [van Rullen and Thorpe, 2001; Oliva and Torralba, In press] to minimize the possibility of feedback, top-down effects in the task. Subjects were asked to respond as fast as they could to the presence or absence of an animal in the image by pressing either of two keys, see Fig. 3.3.

3.5 Results

Performance of the model and of human subjects is similar, even in their pattern of errors (see Fig. 3.4). As for the model, human subject performed best on "close-body" views and worst on "far-body" views. An intermediate level of performance was obtained for "head" and "medium-far" views. Overall no significant difference was found between the level of performance of the model and human subjects. To control for the difficulty of the task, we ran several benchmark systems that rely on simple low-level features (such as simple oriented filters or texture descriptors). Their level of performance was significantly lower (see [Serre et al., 2005a]). Not only the level of performance achieved by the model and human subjects was very similar, but we also found a high correlation between their errors (see [Serre et al., 2005a]). This is consistent with the claim that human subjects and the model use similar strategies.

The success of a feedforward model of the ventral stream - faithful to known quantitative measurements in anatomy and physiology - in achieving human performance in rapid categorization task, and in mimicking its mistakes, suggests that *immediate recognition* – or "vision at a glance" [Hochstein and Ahissar, 2002] - is mainly feedforward over times of about 150 milliseconds from onset of the stimulus. It also raises the hope that we may have the skeleton of a computational and functional model of the ventral stream of visual cortex upon which we may now begin to graft many more detailed computational roles for specific visual areas and neurons.

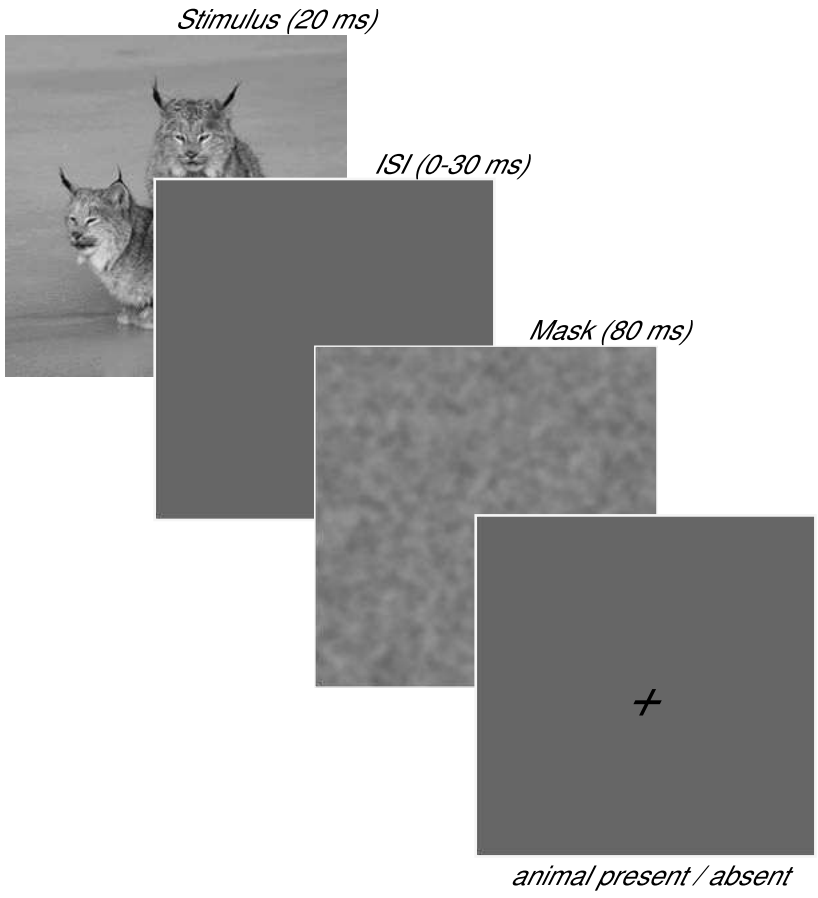


Figure 3.3: Testing humans on an ultra-rapid animal *vs.* non-animal categorization task, modified from [Thorpe et al., 1996]. Each image was shown for 20 ms, followed by a mask appearing after a delay of 30 ms (1/f noise). At the end of the trial, subjects had to press either of two keys for animal present or absent.

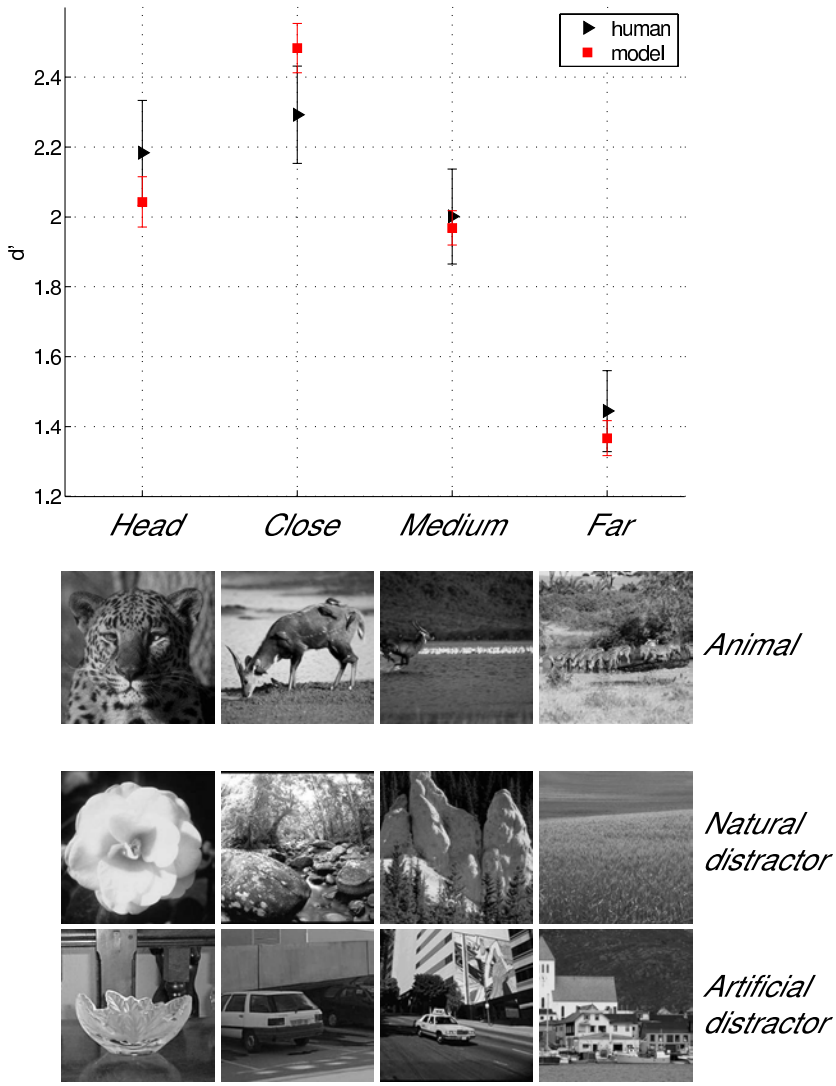


Figure 3.4: Results obtained on the four animal categories (with matching distractors) for both humans and the model. The performance is measured by the d' value, which combines for each subject (or each run of the model on random splits), the hit and false-alarm rates into a single standardized score [Macmillan and Creelman, 1991, see]). Error bars indicate the standard error but there is no direct relation between errors as computed for the model and as computed for human subjects. There is no significant difference in terms of performance between the model and human subjects. Inset: Examples from the four image categories.

Notes

[3]The performance comparison between humans and the model was done in collaboration with Aude Oliva.

4 Visual areas

Predictions made by the original model The model interprets several existing data from system physiology to cognitive science. The original model [Riesenhuber and Poggio, 1999b] made also a few predictions ranging from biophysics to psychophysics. Table 3 summarizes the main model predictions.

Max operation in cortex	The model predicted the existence of complex cells in V1 [Lampl et al., 2004] and V4 [Gawne and Martin, 2002] performing a softmax pooling operation
Tolerance to eye movements	From the softmax operation – originally introduced to explain invariance to translation in IT – the model predicts stability of complex cells responses relative to small eye motions
Tuning properties of view-tuned units in IT	The model has been able to duplicate quantitatively the generalization properties of IT neurons that remain highly selective for particular objects, while being invariant to some transformations [Logothetis et al., 1995; Riesenhuber and Poggio, 1999b] their tuning for pseudo-mirror views and generalization over contrast reversal. Also, the model qualitatively accounts for IT neurons responses to altered stimuli [Riesenhuber and Poggio, 1999b], *i.e.*, scrambling [Vogels, 1999], presence of distractors within units receptive fields [Sato, 1989] and clutter [Missal et al., 1997]
Role of IT and PFC in categorization tasks	After training monkeys to categorize between "cats" and "dogs", we found that the ITC seems more involved in the analysis of currently viewed shapes, whereas the PFC showed stronger category signals, memory effects, and a greater tendency to encode information in terms of its behavioral meaning [Freedman et al., 2002] (see also subsection 4.4)
Learned model C2 units compatible with V4 data	We have recently shown (see Subsection 4.2) that C2 units that were passively learned from natural images seem consistent with V4 data, including tuning for boundary conformations [Pasupathy and Connor, 2001], two-spot interactions[Freiwald et al., 2005], gratings [Gallant et al., 1996], as well as the biased-competition model [Reynolds et al., 1999]
"Face inversion" effect	The model has helped [Riesenhuber et al., 2004] guide control conditions in psychophysical experiments to show that an effect that appeared to be incompatible with the model turned out to be an artifact

Table 3: Some of the correct predictions by the model

4.1 V1 and V2

4.1.1 V1

Our theory of the tuning operation assumes that tuning of simple cells in V1 is due jointly to the geometry of the LGN inputs – corresponding to the non-zero synaptic weights w_j in Eq. 1 – and to intracortical lateral, possibly recurrent, inhibition implementing the normalization operation of Eq. 1. The activity of each simple cell is normalized by the activity of its inputs (from the LGN via interneurons in the feedforward case of Eq. 1 or from other simple cells, mostly with different orientations in the recurrent case) over a large neighborhood (possibly corresponding to the "non-classical" receptive field), since the normalization pool should include LGN inputs which feed simple cells with other orientations.[1]

In the current model implementation all stages from the retina to V1 are modeled in a single step, *i.e.,* simple cell like responses are obtained by directly filtering the input image with an array of Gabor filters at four different orientations, different sizes and all positions (see Fig. 4.1). Also note that in the present version of the model (unlike the original one [Riesenhuber and Poggio, 1999b]) all the V1 parameters are derived exclusively from available V1 data and do not depend – as they did in part in the original HMAX model – from the requirement of fitting the benchmark paperclip recognition experiments (see [Serre et al., 2004] for a comparison between the new and the old model parameters). Thus the fitting of those paperclip data by the new model is even more remarkable than in the original HMAX case. In Appendix A.1 we describe how S1 and C1 parameters were adjusted so that the corresponding units would match the tuning properties of cortical parafoveal cells in response to standard stimuli. In Appendix A.2 we describe how the tuning properties of S1 and C1 units was assessed and give a summary of the S1 and C1 tuning properties (*i.e.,* orientation and frequency tuning). The complete experiments are described in [Serre et al., 2004].

Notes

[1]In our current model implementation this neighborhood is restricted to the classical receptive field. There are many reports of surround suppression or other non-classical receptive field effects in V1 [Cavanaugh et al., 2002]. Although the neural mechanisms for those effects remain controversial (*e.g.,* feedback from higher layer *vs.* lateral interaction), some of them can be explained by a divisive normalization by a large summation field enclosing the classical receptive field. Such mechanism is similar to Eq. 1 and fits well within our framework of the model. Hence, it would be possible to incorporate some of the non-classical receptive field effects by enlarging the neighborhood for the normalization.

[2]It is interesting to notice that Eq. 1 is consistent with tuning of simple cells in V1 being due to the LGN inputs and to intracortical inhibition (the denominator term). The tuning Eq. 1 and 2 predict that the intracortical normalization sharpens the tuning induced by the afferent inputs (the non-zero w_j). Thus the overall receptive field of simple cells would depend on both the afferents from the LGN and on the more numerous intracortical inputs. In the equations the inhibition is assumed to be feedforward but recurrent inhibition may work in a similar way, at least in this version of the model which does not have any dynamics (see also section 5).

[3]Our theory suggest – from Eq. 3 describing the soft-max operation – that there exists a subclass of complex cells which are driven by the most active of its simple cells (or simple cells-like) afferents.

[4]Livingstone & Conway [Livingstone and Conway, 2003] reported an interesting pattern of nonlinear interaction within the receptive field of direction-selective complex cells in V1. In Appendix A.7, we show how the converging inputs from simple cells combined with divisive normalization, both of which are consistent with our framework of V1 complex cells, can create similar interaction patterns.

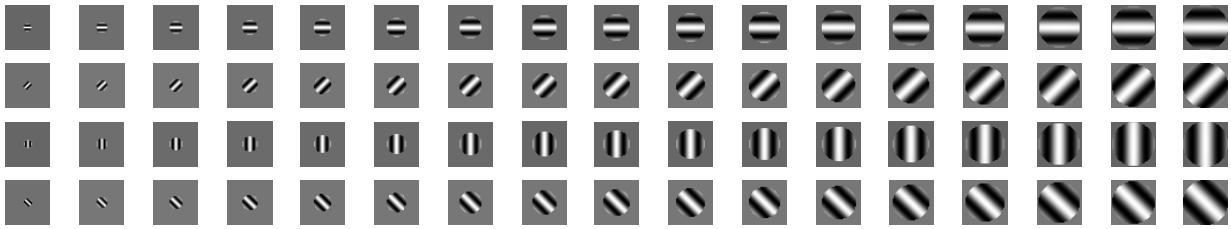


Figure 4.1: The S1 unit receptive fields (Gabor functions). Receptive field sizes range from 0.19 to 1.07 degrees at four different orientations. In order to obtain receptive field sizes within the bulk of the simple cell receptive fields ($\approx 0.1 - 1$ degree) reported in [Schiller et al., 1976a; Hubel and Wiesel, 1965b], we cropped the Gabor receptive fields and applied a circular mask so that, for a given parameter set (λ^0, σ^0), the tuning properties of units are independent of their orientations θ^0. Note that the receptive fields were set on a gray background for display so as to preserve relative sizes.

4.1.2 V2

In the present implementation of the model we do not distinguish between V1 and V2, effectively assuming that V2 is equivalent to V1. It is however possible that V2 represents an additional stage of simple and complex cells intermediate between V1 and V4 (which are not modeled at present). It is also possible that the S2 cells of the model are mostly in V2 instead than V4 (whereas the C2 cells would probably be in V4, according to the little physiological evidence on invariance of V4 responses to position).

Notes

[5]A max operation could explain in part the stability against small eye movements (say 30 minutes of arc or so, found in some complex cells (at various stages of the visual system) by Poggio and Motter (see Poggio AI Working Paper 258, "Routing Thoughts", 1984).

[6]In the present implementation of the model the tuning of the simple cells in V1 is hardwired. It is likely that it could be determined th¡rough the same passive learning mechanisms postulated for the S2 cells (possibly in V4 and PIT), possibly with a slower time scale and constrained to LGN center-surround subunits. We would expect the automatic learning from natural images mostly of oriented receptive fields but also of more complex ones, including end-stopping units (as reported for instance in [DeAngelis et al., 1992] in layer 6 of V1).

[7]There is evidence for mechanisms in V1 and V2 which may support computations of Gestalt-like properties such as collinearity, filling-in (as demonstrated by illusory contours [Bakin et al., 2000]) and border ownership (see [Zhou et al., 2000]). How to extend the model to account for these properties and how to use them in recognition tasks is an interesting and still open question. Unlike the end-stopping units, these properties may require feedback from higher areas such as V4. In future work we plan to test the model in terms of recognition of figures based on illusory contours such as Kanisza triangle.

4.2 V4

4.2.1 Properties of V4

Situated between V2 and IT, visual area V4 is known to have receptive fields of intermediate sizes (larger than V1 and smaller than IT on average), tuning to features of intermediate complexity (more complex than V1 and simpler than IT), and invariance to small translations [Kobatake and Tanaka, 1994]. Although first known for their color selectivity, neurons in V4 are selective to a wide variety of forms and shapes, such as bars, gratings, angles, closed contour features, and sparse noise, *etc.* [Desimone et al., 1985; Gallant et al., 1996; Pasupathy and Connor, 1999, 2001; Freiwald et al., 2005]. In addition, the selectivity of most V4 neurons described so far is invariant to small translations of about 25% of the receptive field. In this section, we provide some supporting evidences for our theory by modeling individual V4 responses, predicting responses to novel stimuli, and showing that learning from natural images produces a population of model units with compatible statistics measured in several independent V4 experiments. Further results and supporting simulations can be found in Appendix A.8 and in [Cadieu, 2005].

The model assumes that there are *simple* (S2) and *complex* (C2) computational units which differ in their translational and scale invariance properties. The available data from V4 suggests that most of the reported recordings are from C2-like cells (cells with a range of translation invariance that cannot be attributed to the range of invariance from V1 complex cells). The model predicts the existence of S2-like cells. They may at least in part be present in area V2 and feed directly to the C2-like cells in area V4. We do not think there is enough evidence so far for ruling out the presence of *simple* and *complex* cells in V4 (the difference would be mostly in the larger range of invariance to position and scale for C2 cells than S2 cells).

4.2.2 Modeling Individual V4 Responses

Selectivity The model is capable of reproducing V4 selectivity to both grating stimuli and boundary conformation stimuli independently. Eight response sets for V4 neurons (4 measured with the 366 boundary conformation stimuli, referred to as B1, B2, B3, and B4; and 4 measured with the 320 grating stimuli, referred to as G1, G2, G3, and G4) were used to fit model units (C2 units). The result of fitting of neuron B2 is shown in Fig. 4.2, and for G1 is shown in Fig. 4.3. In both figures the response of the V4 neuron and the model unit are plotted side by side for comparison. The magnitudes of those responses are displayed by the gray level of each stimulus. Black indicates a high response, and light gray indicates a low response with intermediate responses mapped linearly to intermediate shades of gray.

Both Fig. 4.2 and Fig. 4.3 show a high degree of correspondences between the neural and the model responses. Stimuli that elicit high responses from the neuron also produce high responses in the model. The degree of sparseness, or sharpness in selectivity, of these neurons is also captured in the model responses. Similar results were achieved for the other 6 V4 neurons examined. Thus, selectivity that has previously been described as tuning for concave or convex boundary elements or tuning for Cartesian, polar or hyperbolic gratings are also reproducible with the model. See Appendix A.8 for the fitting procedure of the V4 neural responses using the model.

Invariance While maintaining selective responses, model unit are also capable of reproducing invariance properties of V4 neurons, on the currently available data set measuring featural and translational invariance.

- **Translation:** Fig. 4.4 shows the responses to a preferred and non-preferred stimuli for V4 neuron B3, adapted from Fig. 6A of [Pasupathy and Connor, 2001], and a model unit (C2 unit) over a 5×5 translation grid. The model unit yields high responses to the preferred stimuli over a translation range comparable to the the V4 cell. For the non-preferred stimuli, the model unit shows low responses over all translated positions. Hence, stimulus selectivity is preserved over translation for the model unit, and the degree of translation invariance is similar to that of the V4 neuron. All model units examined exhibit similar degrees of translation invariance.

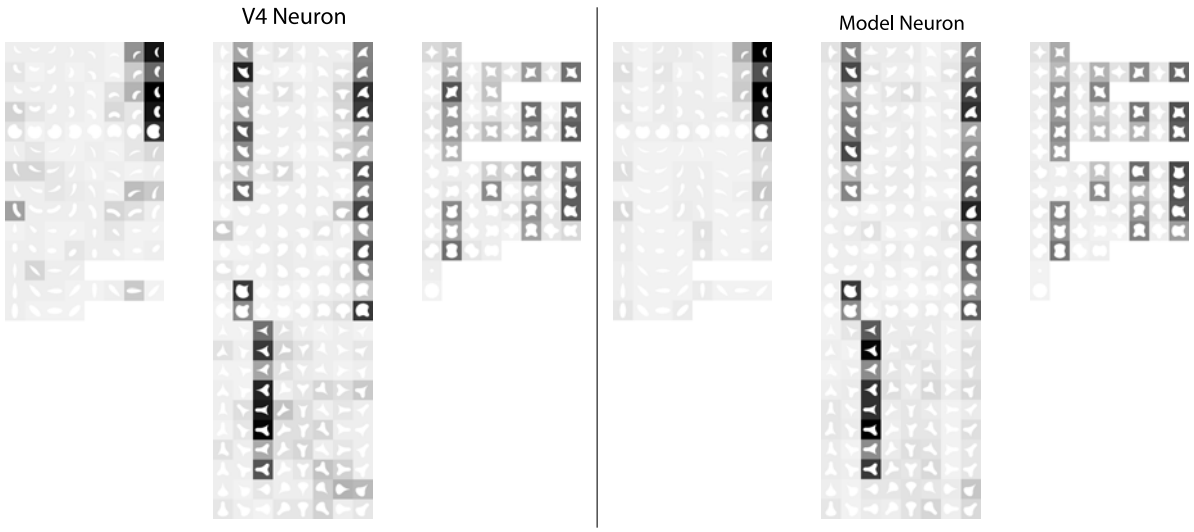


Figure 4.2: V4 neuron B2 (left) and model unit (right) responses over the boundary conformation stimulus set. The response magnitude to each stimulus is displayed by the gray level of the stimulus. Dark, or black, shading indicates a strong response, and light gray indicates a low response. Therefore, the stimuli that are clearly visible are those that elicit the highest responses. The V4 neuron response has been adapted from Fig. 4 of [Pasupathy and Connor, 2001].

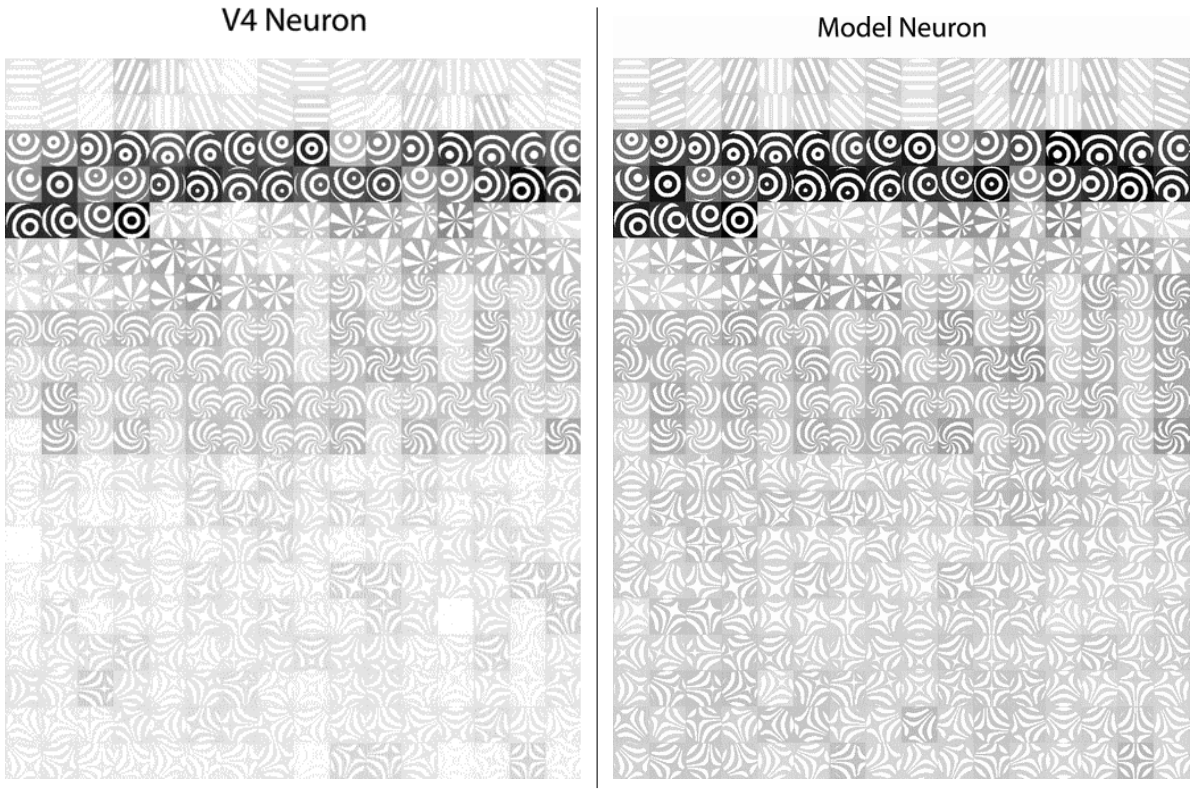


Figure 4.3: V4 neuron G1 (left) and model unit (right) responses to the grating stimulus set. Display conventions are the same as in Fig. 4.2. V4 data is courtesy of Freiwald, Tsao, and Livingstone [Freiwald et al., 2005].

While S2 units in the model demonstrate selectivity that is dependent on absolute spatial position (with the exception of small invariance properties inherited from the C1 representations), C2 units demonstrate selectivity that is independent of absolute spatial position. C2 units inherit the selectivity characteristics of their afferent S2 units, but achieve translation invariance by pooling (with max or softmax operation) over S2 units that cover different parts of receptive fields. As a result, C2 units transform absolute position tuning of the S2 units into relative position tuning, typical of V4 responses.

- **Scale:** There is currently a lack of experimental data indicating the degree of scale invariance in area V4. Many models of visual cortex, including ours, postulate increasing scale invariance along the ventral pathway; such measurements should therefore provide important constraints for such models.

- **Featural:** Fig. 4.2.2 shows the responses of a V4 neuron and a model unit fit to the V4 neuron's response over the stimulus set adapted from Fig. 6B of [Pasupathy and Connor, 2001], and the model unit shows a response pattern that is nearly identical to the V4 neuron's response. This particular model unit has 8 C1 subunits, although similar tuning can be seen with fewer subunits. The model response is highly correlated with the angular position of the convex extremity and poorly correlated with the orthogonal offset of the extremity. This result matches 26/29 V4 neurons measured in [Pasupathy and Connor, 2001].

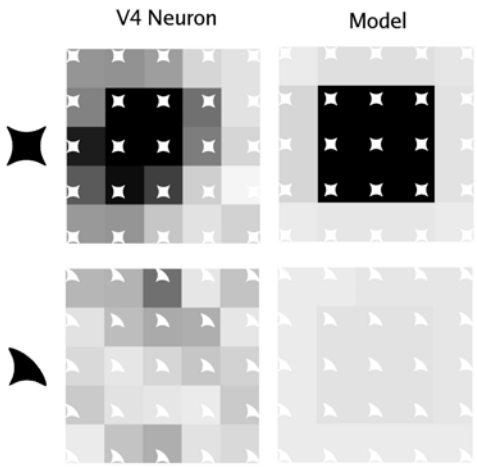


Figure 4.4: V4 neuron B3 and model unit responses to translated stimuli. Both show comparable translation invariance. V4 neuron response is adapted from Fig. 6A of [Pasupathy and Connor, 2001].

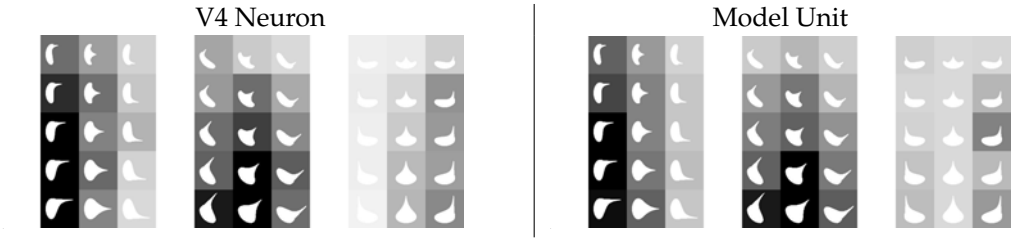


Figure 4.5: V4 neuron B3 and model response to the relative position boundary conformation stimulus set. The model unit was fit to the V4 neuron's response pattern on this stimulus set (45 stimuli). The V4 response is adapted from Fig. 6B of [Pasupathy and Connor, 2001].

4.2.3 Predicting V4 Response

The fitting procedure described in Appendix A.8 is used to predict responses to the stimuli of the same type, either boundary conformation or grating stimuli, and to predict responses to a novel stimulus class, 2-spot stimuli, from grating stimulus responses. Results are presented here for 8 neurons for within-class predictions: 4 measured with a boundary conformation stimulus set and 4 measured with a grating stimulus set. The 4 neurons measured with the grating stimulus set were also measured with 2-spot stimuli allowing predictions across stimulus classes.

Within-Class Predictions For within stimulus class predictions, neuron responses were predicted using a cross-validation methodology with 6 folds (61 stimuli per fold) for the boundary conformation stimulus set and 4 folds (80 stimuli per fold) for the grating stimulus set. The cross-validation procedure predicts the responses in each fold of the data. For each fold, the forward selection algorithm terminates when the mean squared error on the training set improves by less than 1%. The model unit is then used to predict the responses for the test set. The predictions, combined for each fold, span the entire dataset. Therefore, we have a prediction for each data point within the stimulus set. This produces within-class predictions (the model predicts responses to stimuli of the same class as used to fit the model). The results for within-class predictions are summarized in Table 4.

Neuron Label	Correlation			Num. of Subunits
	Train	Test	Pred.	
B1	0.86	0.78	0.77	17.0
B2	0.94	0.85	0.86	21.0
B3	0.90	0.77	0.76	19.8
B4	0.82	0.70	0.71	14.7
G1	0.95	0.88	0.87	22.3
G2	0.80	0.70	0.68	15.3
G3	0.88	0.76	0.76	18.0
G4	0.88	0.73	0.76	16.8
Mean	0.88	0.77	0.77	18.0

Table 4: Results of fitting model units to V4 neurons. Neuron label indicates the name of the V4 neuron used (B for Boundary conformation stimulus set and G for grating stimulus set). For each fold of the data a correlation coefficient (Correlation) is found for the training set and the test set. These values are then averaged over the folds. The prediction column (Pred.) indicates the correlation coefficient between the union of the test set predictions and the V4 response. The final column indicates the average number of subunits used in each model averaged over the folds (the algorithm stopped when the MSE (mean squared error) on the training set did not improve by more than 1%). The bottom row shows the average of the various results over all 8 V4 neurons.

Predicting 2-Spot Interactions A novel approach to investigating V4 representation and computation has been the application of 2-spot reverse correlation. This technique has been used in V1 and MT to determine the substructure of receptive fields [Livingstone and Conway, 2003; Szulborski and Palmer, 1990] (also see Appendix A.7). The reverse correlation mapping procedure calculates the interaction produced by two stimuli as a function of their relative position. Such a sampling may be less biased than the use of gratings or contour stimuli. Data from V4 by Freiwald, Tsao and Livingstone show an intricate pattern of interactions within the receptive field [Freiwald et al., 2005]. Such a technique is a useful additional tool for understanding the possible computations a neuron is performing on its inputs.

In this section we examine four V4 neurons that have been measured using grating stimuli and have also been measured with a 2-spot reverse correlation technique by Freiwald and colleagues. The model, because it is fully quantitative, is capable of generating responses to both of these stimulus classes. Model parameters can be fit to a V4 neuron's grating stimulus response and then used to predict the response of the neuron to the 2-spot reverse correlation technique. For an analysis of the factors that produce 2-spot interaction in model units see Appendix A.8.2.

The results of fitting model parameters to the grating responses of four V4 neurons and then predicting the 2-spot interaction maps are shown in Fig. 4.6. Further details on the methodology used to predict 2-spot interaction maps can be found in Appendix A.8. Each interaction map is displayed so that the origin (the center of each image) corresponds to 2-spots occurring at the same point. The color map indicates facilitation, red, inhibition, blue, and balanced or no effect, green. The predictions for three of the V4 neurons, G1, G2, and G4, display qualitative agreement with the measured 2-spot reverse correlation maps. For example, neuron G1, in the first row, displays a wreath like facilitatory interaction and a gap or depression in the interaction along the $135°$-$315°$ axis. The prediction and measurement for neuron G2 show a strong facilitatory interaction along the $135°$-$315°$ axis. Similarly, the prediction and measurement for neuron G4 show strong facilitatory interaction along the $90°$-$270°$ axis. However, there is no clear correspondence between the prediction and measurement for neuron G3.

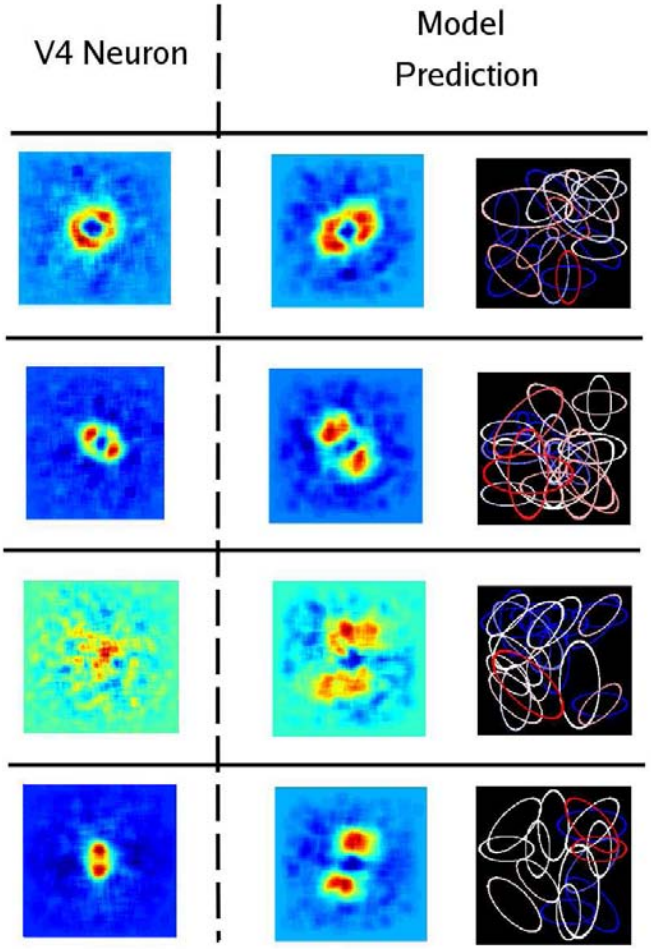


Figure 4.6: Predicted 2-spot interaction maps for 4 V4 neurons. Rows correspond to the results for each of the 4 neurons, labeled as: G1, G2, G3, and G4. From left to right, the columns indicate experimental 2-spot interaction map, the C2 unit's predicted 2-spot interaction map, and a visualization of S2 unit selectivity. The visualization shows C1 subunits as colored eclipses, which feed into the S2 units that then feed into the C2 unit used to model the V4 grating response and predict the 2-spot interaction map. The predictions for neurons G1, G2, and G4 qualitatively match the experimental measurements. However, the prediction for neuron G3 does not match the measurement. V4 data is courtesy of Freiwald, Tsao, and Livingstone [Freiwald et al., 2005].

4.2.4 Model Units Learned from Natural Images are Compatible with V4

In the previous sections, we showed that model units could fit properties of V4 cells and predict their responses to different stimuli. In this section, we show that the unsupervised learning stage described in Section 2.2 generates S2 and C2 units with properties consistent with physiological data. In other words, selectivity in area V4 may be learned through passive viewing of the environment. By learning activity patterns during the presentation of natural images, model units demonstrate selectivity that is characteristic of V4 neurons under three different experimental conditions: selectivity to boundary conformation [Pasupathy and Connor, 2001], selectivity to grating class [Gallant et al., 1996], and two stimuli interactions in the absence of attention [Reynolds et al., 1999].

- **Boundary Conformation Stimuli:** Fig. 4.7 displays the results of the simulated experimental methodology from [Pasupathy and Connor, 2001] performed on a population of 109 model units. Each model unit has learned the pattern of input from a natural image patch. The resulting population of model units exhibits higher tuning in boundary conformation space than edge orientation tuning space or axial orientation space, a characteristic of V4 neural population as reported in the experiment.[9]

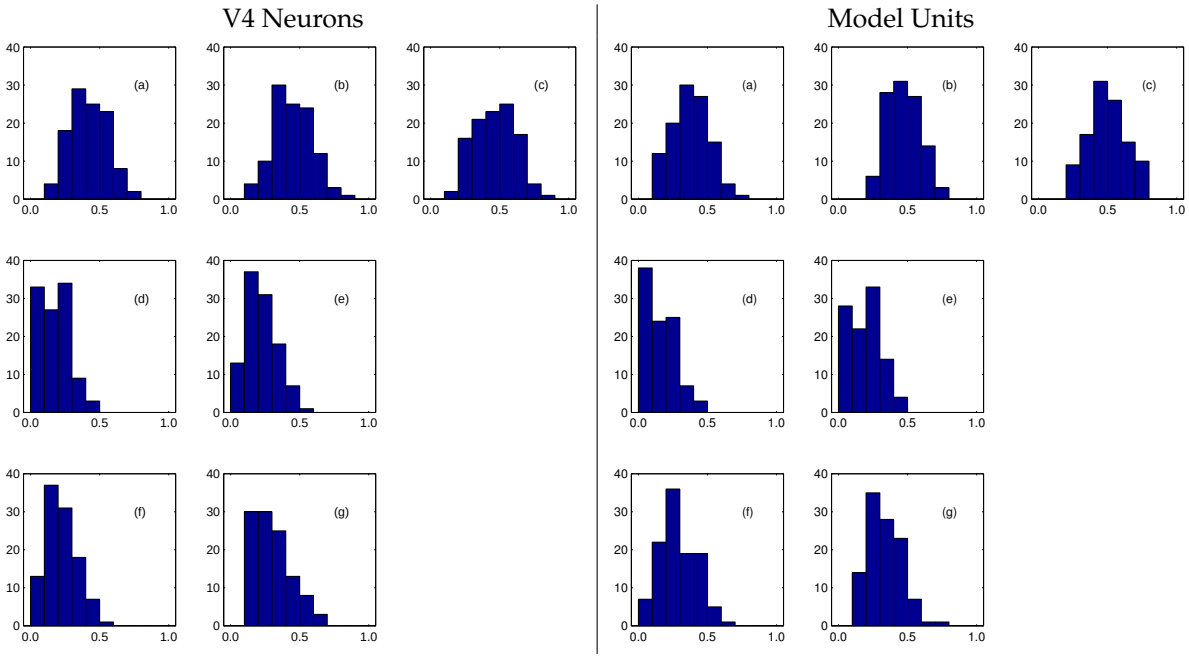


Figure 4.7: A population of model V4 neurons learned from natural images displays tuning properties consistent and comparable with the experimental results using boundary conformation stimuli [Pasupathy and Connor, 2001]. The selectivity of 109 model units has been learned according to the methodology described in Section 2.2. The seven panels (on left from the original experiment and on right from the model) display the population histogram of the correlation coefficient for a different tuning function. (a) 2D boundary conformation, (b) 4D boundary conformation, (c) 2-Gaussian 4D boundary conformation, (d) edge orientation, (e) edge orientation and contrast polarity, (f) 2-D axial orientation elongation tuning functions, and (g) 3-D axial orientation$\times$length$\times$width tuning functions. V4 neurons characteristically show higher correlation coefficients for boundary conformation tuning functions (a, b, c) than for edge orientation or axial orientation tuning functions (d, e, f, g), and the model units show the same trends. The median correlation coefficients (for figures a through g) for the V4 data are (0.41, 0.46, 0.46, 0.15, 0.21, 0.18, 0.28), whereas for the learned model units are (0.38, 0.47, 0.50, 0.11, 0.18, 0.28, 0.32). The V4 data is courtesy of Pasupathy and Connor.

- **Grating Stimuli:** Fig. 4.8 displays the grating class selectivity of a population of 109 model units with selectivity learned from natural images. Within each panel, the highest response to polar, hyperbolic, or Cartesian stimuli for each V4 neuron (experimental or model) is plotted in a 3-dimensional space. V4 neurons characteristically show higher relative response to hyperbolic and polar gratings than Cartesian gratings (with slightly stronger bias toward polar gratings). See [Kouh and Riesenhuber, 2003] for an analysis using the old, hard-wired features.

- **Two Bar Stimuli:** Reynolds, Chelazzi and Desimone measured the responses of V4 neurons to single or two oriented bars with or without attention [Reynolds et al., 1999]. Their experiments have found that in the absence of attention (as is the case in the model), the addition of a second stimulus presented within the receptive field of a V4 neuron causes the neuron's response to move toward the response of the second stimulus alone. To test this result in the model, a reference stimulus and a probe stimulus, each an oriented bar, are presented to model units individually or at the same time. The responses are then analyzed using the same methodology as in the original experiment.[10] The results of this experiment are shown in Fig. 4.9. The experimental findings are reproduced on the left, and the model results are on the right.[11]

Notes

[8]The authors would like to thank A. Pasupathy and C. E. Connor for their help in reconstructing their boundary conformation stimuli and analysis methods, as well as the data for the 109 neurons in [Pasupathy and Connor, 2001], used in Fig. 4.7. The authors also thank W. Freiwald, D. Tsao, and M. Livingstone for providing neural data to the grating and 2-spot stimuli and for many helpful discussions [Freiwald et al., 2005].

[9]The correlation coefficients in Fig. 4.7 were found using the same nonlinear fitting procedures with different tuning functions as described in Fig. 9 of [Pasupathy and Connor, 2001]. There are some small numerical differences between our results and those of Pasupathy and Connor. The discrepancies may be due to differences in conventions for extracting parameters (curvature, edge orientation, axial orientation) from the stimuli, and differences in the nonlinear fitting routines (*e.g.*, number of initial points).

[10]There is a slight change in the "experiment": in the original experimental method, the bar stimuli are parameterized by orientation and color, while in our simulations, the color dimension is replaced by increasing the number of sampled orientations, because the model currently contains no color information. This results in an equal number of measurements for each neuron under the experimental recordings and the simulation.

[11]Interestingly, such interference effects (*i.e.*, presenting preferred and non-preferred stimuli together produces the response somewhere between the individual responses, sometimes close to an average, in the absence of attention) may be occurring in other cortical areas, as shown in Section 4.3. The model accounts the clutter effect *regardless of any particular cortical area*, using the same principle operations for selectivity and invariance appearing across different layers. In fact, the biased competition model devised to explain the results of [Reynolds et al., 1999] is closely related to Eq. 1 in our model. Since normal vision operates with many objects appearing within the same receptive fields and embedded in complex textures (unlike the artificial experimental setups), understanding the behavior of neurons under such clutter condition is important and warrants more experiments.

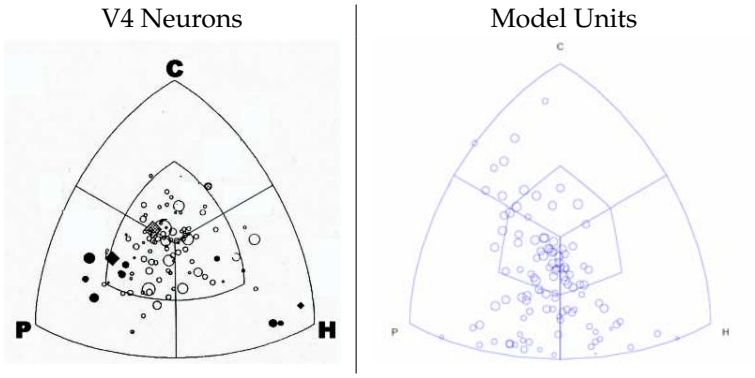


Figure 4.8: A population of model V4 neurons learned from natural images displays tuning properties characteristic of experimental results in V4 to a set of polar, hyperbolic, and Cartesian grating stimuli. The selectivity of 109 model units was learned according to the methodology described in Section 2.2. The same experimental approach is simulated for the model units learned from natural images. The panel on the left displays the experimental findings for 103 V4 neurons (adopted from Fig. 4a in [Gallant et al., 1996]) and the panel on the right displays the results for the model units. Within each panel, the highest response to polar, hyperbolic, or Cartesian stimuli for each V4 neuron (experimental or model) is plotted in a 3-dimensional space. V4 neurons characteristically show higher relative response to hyperbolic and polar gratings than Cartesian gratings (with slightly stronger bias toward polar gratings). See [Kouh and Riesenhuber, 2003] for an analysis using the old, hard-wired features.

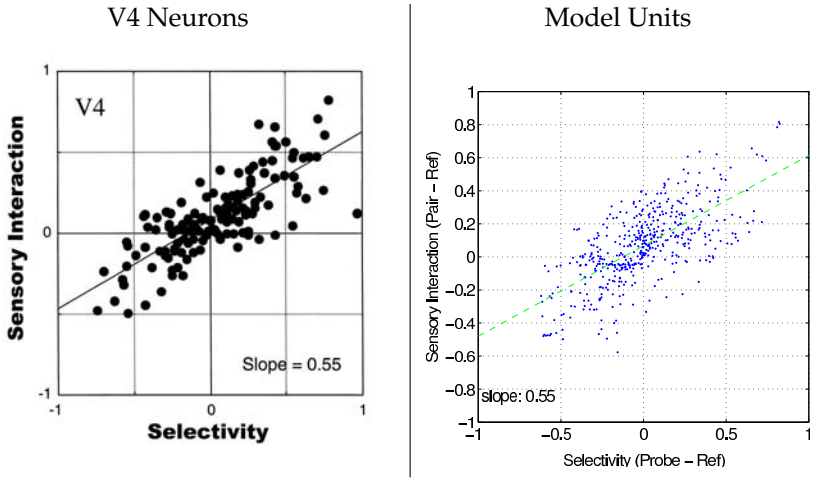


Figure 4.9: The model exhibits similar behaviors to the two bar stimuli as the V4 neurons in the absence of attention. The summary of V4 neural responses, adopted from Fig. 5 in [Reynolds et al., 1999], is shown on the left. The addition of a stimulus moves the response toward the response to that stimulus alone, or, put differently, the response to the clutter condition lies between the responses to the individual stimulus.

4.3 IT

4.3.1 Paperclip experiments

The first version of the model [Riesenhuber and Poggio, 1999b] was originally motivated by the experiments in [Logothetis et al., 1995], where monkeys were extensively trained to recognize a set of novel "paperclip" objects and some neurons in anterior IT were found to be tuned to the trained views of those objects in a scale, translation, and 3D rotation invariant manner. In other words, these view-tuned neurons responded more strongly to the scaled, translated, and rotated (in depth) images of the preferred paperclip than to a large number (60) of distractor paperclips, *even though these objects had been previously presented at just one scale, position and viewpoint*.

Of fundamental importance in this experimental design is the use 1) of novel object class that monkeys had not had any visual experience and 2) of distractor objects, which allows one to define the range of selective invariance. Measuring how much the response changes to the object transformation is not as meaningful as making a comparison to a threshold or a reference value. Using the maximum response to many distractor objects as a threshold, one can define the range of invariance to the transformed versions of the preferred object.

The model of [Poggio and Edelman, 1990] proposed how view-tuned cells can produce view-invariant cells, while not addressing how such view-tuned cells could come about. The theory described here shows that the hierarchical, feedforward architecture with biologically plausible (and necessary for object recognition) operations, as shown in Fig. 2.1, can account for the data of [Logothetis et al., 1995], which is regarded as our main benchmark test. We emphasize that the model parameters are tuned to fit the properties of the lower layers (corresponding to V1 and V4), not the higher layer (corresponding to IT), yet, by tuning to just one view of the target object, the model shows similar ranges of viewpoint rotation and scale invariance as the IT neurons.

Fig. 4.10 shows the response of one model unit from the S4 layer, also known as VTU (view-tuned unit), which was created by tuning it to a particular view of a paperclip stimulus. As the object is slowly transformed away from the preferred view, via 3D rotation, scale or translation, the response of this model unit decreases. However, over certain ranges of transformations, the magnitude of its response is greater than to different paperclips. This particular model unit shows the rotational invariance range of 32°, scale invariance range of 2.7 octaves, and translation invariance range of 4° of visual angle. Other model units tuned to other reference paperclips show similar range of invariances, in good agreement with the experimentally observed values [Logothetis et al., 1995].[12]

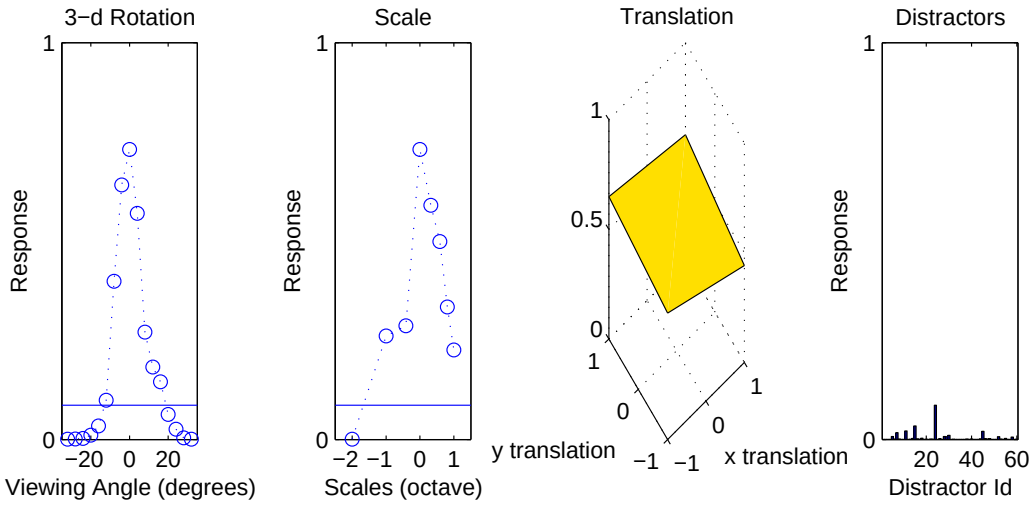


Figure 4.10: Paperclip benchmark test results: For rotation experiments, viewpoints were varied in 4 degree steps. For scale, the stimulus size varied from 16 to 128 pixels in 16 pixel steps. The axes in the translation figure are in units of $\times$ 2 degrees of visual angle. In this simulation, softmax (for C1 and C2b units) and normalized dot product with sigmoid (for S2b and VTU) were used on features learned from paperclip images.

Notes

[12]The first version of the model [Riesenhuber and Poggio, 1999b] had already shown the agreement with the paperclip experiments. Here we presented simulation results using the extended version of the model – with normalized weighted sum operations and learned features (from paperclip images), as described in previous sections. The results are consistent with the original version of the model.

[13]Paperclip stimuli were also used in [Bülthoff and Edelman, 1992] for psychophysics and in [Poggio and Edelman, 1990] for modeling.

[14]A more recent study [Hung et al., 2005a] confirms the scale and translation invariance of IT neurons using different novel objects. See Section 4.3.3 on the readout experiments.

4.3.2 Multiple object experiments

The paperclip experiments mentioned above showed that IT neurons (and the model units) have transformation (scale, translation, and 3D rotation) invariant representation of an object, necessary for a robust object recognition performance. Another challenging condition, under which the visual system continues to perform robustly, is when the scene is cluttered with multiple objects or the object is embedded in a complex background (an extreme clutter condition).

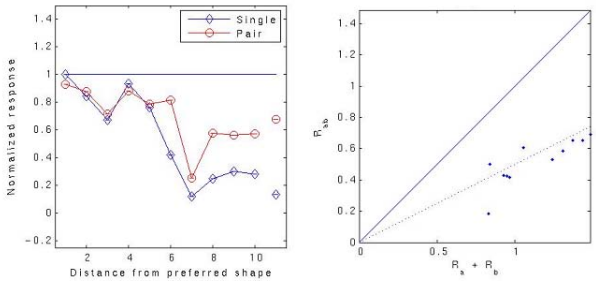
Results from physiology and comparison with the model A recent study [Zoccolan et al., 2005] systematically looked at the effect of adding one or two distractor object within the receptive field of an IT neuron. Interestingly, most recorded neurons (which was probably a small and non-random fraction of visually responsive neurons[16]) showed an average-like behavior. That is, the response to the cluttered condition, containing two or three objects, was close to an average of responses to the individual objects presented alone. Here we report some agreements and disagreements between the experiment and the simulation results.

Agreements: As shown in Fig. 4.11, some model units exhibit similar multiple object effects when tested with the same stimuli used in Experiment 1 and 2 of [Zoccolan et al., 2005]. For the top figure, a view-tuned unit (analogous to an IT neuron, see Fig. 2.1) is created so that its optimal stimulus is a face image presented alone. Therefore, when other face images, morphed smoothly away from the optimal stimulus, are presented, the response decays gradually. When the optimal and the non-optimal images are presented together, the response of this model neuron falls between the responses to the individual stimuli, showing an average effect. As shown in Fig. 4.11c, the average effect is also observed with three stimuli (star, triangle, and cross shapes), although there seems to be a bigger spread.

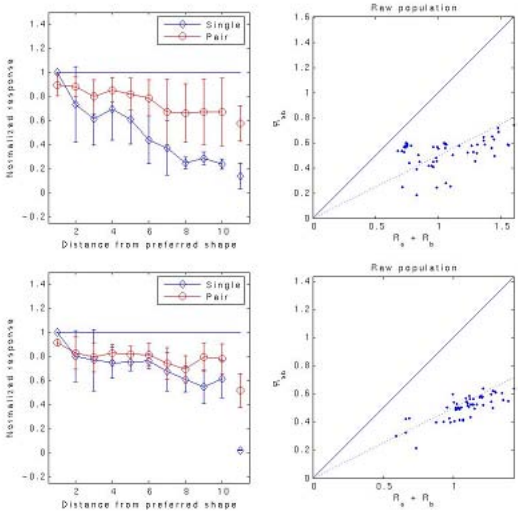
Disagreements: We emphasize that not all model units behave according to the "average rule." Some units are more invariant to the clutter (response to the clutter stays closer to the optimal), and others are not (response is degraded by the non-optimal cluttering stimuli). There is a wide variety of response patterns to the clutter condition than just the "averages" reported in [Zoccolan et al., 2005]. However, using different selection criteria for the IT neurons, Zoccolan and colleagues are currently discovering many more "non-average" neurons, and the preliminary analysis shows an interesting relationship between the selectivity of the neuron and the average-like behavior, which may account for the discrepancy between the experiment and the simulation results. [Zoccolan etal., personal communication] We are in close collaboration to understand this effect, and it may result in a better agreement with respect to the fraction of neurons with clutter tolerance or intolerance.

Possible explanations for the "average" behavior For those particular neurons showing an "average" response to the clutter condition, what are the underlying mechanisms for such behavior? There are several potential explanations, many of which were already listed in [Zoccolan et al., 2005]. A difficult and unlikely explanation is to assume a separate mechanism for counting the number of objects present in the receptive field and forming a divisive factor.[17] Another explanation is to assume a normalization stage by the overall activation of the entire IT population. This is a quite feasible, maybe even desirable, mechanism, since such normalization would make learning easier for the next layer (*e.g.*, PFC, see Fig. A.10 in Appendix A.6.) However, here we explore two explanations directly provided by the theory.

(a) Experiment using morphed face images (single unit)



(b) Experiment using morphed face and car images (averages)



(c) Experiment using star, triangle, cross stimuli

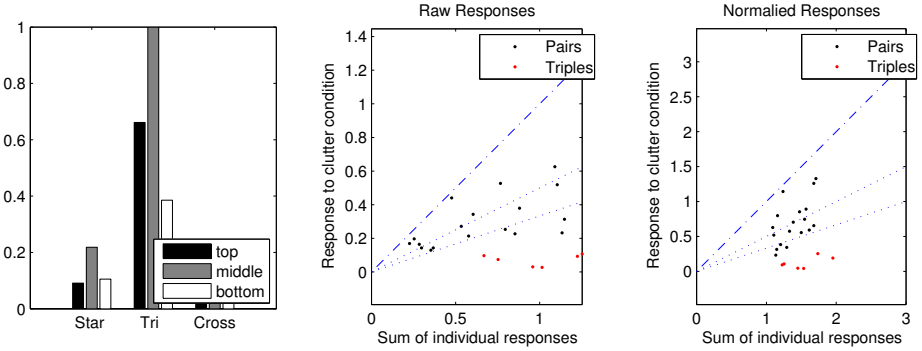


Figure 4.11: (a) These figures show the response of view-tuned units (analogous to IT neurons) to the single and clutter stimuli, tuned to one end of the morphline. The rightmost point in the first column is using the optimal stimulus and non-optimal stimulus from a different stimulus class. For example, in study with face morphlines, a car image was used, and vice versa for the car morphlines. (b) Average results from 10 different view-tuned units: 5 units are tuned to face images (top) and the other 5 are tuned to car images (bottom). (c) The first column shows the model unit's responses to the single stimulus condition (at three different positions within the receptive field). There are 100 randomly chosen afferents (C2b units), which are tuned to, or learned from, natural images. This unit is tuned to a triangle shape (plus some mean-zero noise). Second column shows that the responses to the clutter conditions are clustered around the average line for both pairs and triples conditions. The third column shows the results with normalized responses (same as Fig. 6 in the experimental paper). The convention of these figures is the same as Fig. 3, 5 and 6 from [Zoccolan et al., 2005]. In this simulation, softmax (for C1 and C2b units) and normalized dot product with sigmoid (for S2b and VTU) were used on features learned from natural images.

(a) Possible average mechanism (b) Average effect with multiple objects

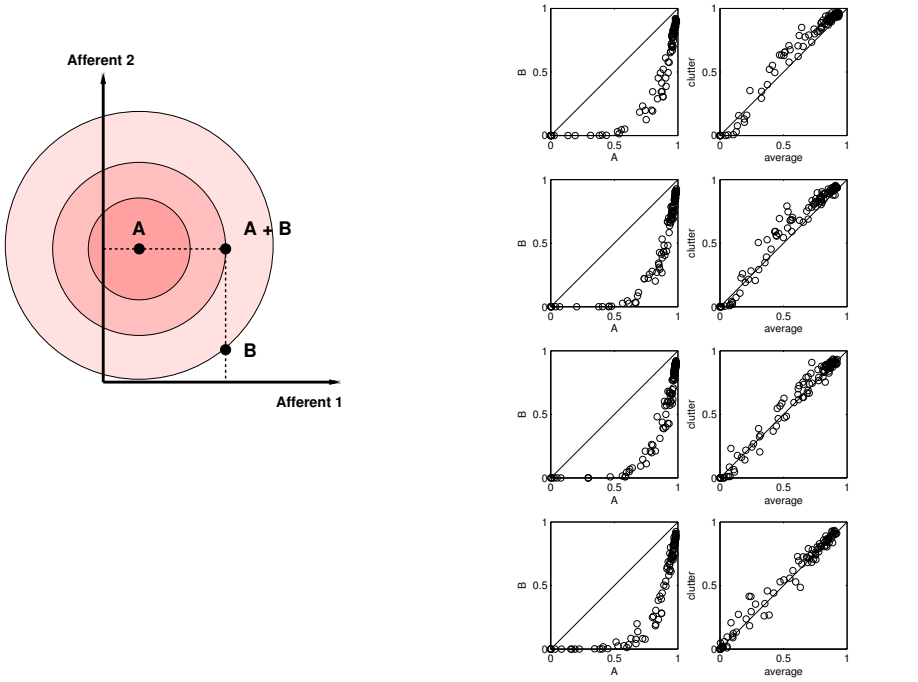


Figure 4.12: (a) One of the mechanisms in the model indicating that the average effect in some cells is produced by the alternation of invariance (max-like) and selectivity (Gaussian-like tuning) operations. In this two dimensional example, suppose that there are two afferent neurons that are activated differently to stimulus $\mathbf{A}$ and $\mathbf{B}$ as indicated. When both stimuli are presented, the response of these afferents are determined by a maximum operation, as indicated by $\mathbf{A} + \mathbf{B}$. Suppose that the response of the neuron under study is determined by the bell-shaped tuning operation over these two afferents, and in fact, tuned to stimulus $\mathbf{A}$. The response to other stimuli, or other patterns of activations, is determined by this tuning operations indicated by the concentric circles around $\mathbf{A}$ in the figure. Note that the combination of these two operations makes the response to the clutter an approximate average of the responses to each stimulus. (b) A toy simulation of the average effect with more than two objects (n = 3, 5, 10, and 100, from top to bottom), using artificially generated activations of 50 afferents. The response of each afferent is randomly generated (using a uniform distribution between 0 and 1) for n individual objects. For the clutter condition when n objects are present together, the response of each afferent is assumed to be the maximum of those n values. The response of the output neuron (receiving inputs from those afferents) is tuned to one of the n objects using a bell-shaped tuning function. For this toy simulation, we created 100 such artificial output neurons. As shown in the first column, the response to the tuned object $\mathbf{A}$ is greater than some other object $\mathbf{B}$. Note that the response to the clutter and the average response to the individual n objects are quite similar. In other words, the alternation of invariance and selectivity operations may produce a response for multiple objects close to the average.

A key mechanism in the model which could be responsible for some average effects is the widespread normalization operation for selectivity and invariance (see Section 2). Consider the following normalized dot product operations (same as Eq. 1), when two stimuli $\mathbf{A}$ and $\mathbf{B}$ are presented together, and note how average-like behavior may result under certain situations.[18]

$$y(\mathbf{A}+\mathbf{B}) = \frac{\sum_{j=1}^{n} w_j \, x_j(\mathbf{A}+\mathbf{B})^p}{k + \left(\sum_{j=1}^{n} x_j(\mathbf{A}+\mathbf{B})^q\right)^r} \simeq \frac{\sum_{j=1}^{n} w_j \, (x_j(\mathbf{A})^p + x_j(\mathbf{B})^p)}{k + \left(\sum_{j=1}^{n} (x_j(\mathbf{A})^q + x_j(\mathbf{B})^q)\right)^r} \tag{8}$$

$$\simeq \frac{\sum_{j=1}^{n} w_j \, x_j(\mathbf{A})^p}{k + 2^r \left(\sum_{j=1}^{n} x_j(\mathbf{A})^q\right)^r} + \frac{\sum_{j=1}^{n} w_j \, x_j(\mathbf{B})^p}{k + 2^r \left(\sum_{j=1}^{n} x_j(\mathbf{B})^q\right)^r} \simeq \frac{y(\mathbf{A}) + y(\mathbf{B})}{2^r}. \tag{9}$$

Note that Reynolds and Desimone's biased competition model in V2 and V4 [Reynolds et al., 1999] uses the same type of equation (normalized dot product) and a similar set of assumptions (two stimuli activate independent population of the afferent neurons, and $p = q = r = 1$). It makes thus sense that this model accounts for the average-like behavior occurring in V2 or V4 cells when two stimuli are presented together (but are not too close).

The theory provides an additional reason for the average effect behavior of some cells: the alternation of the invariance and, then, of the selectivity operation, as schematically illustrated in Fig. 4.12a. Suppose that the stimuli $\mathbf{A}$ and $\mathbf{B}$ activate a set of afferent neurons to different levels. When they are presented together, the response of those afferents is determined by the maximum operation in the *complex* units in the model (*e.g.*, C2 units). Now, suppose the cell being recorded is a *simple* cell receiving inputs from those afferents and tuned to a particular pattern generated by one of the stimulus (*e.g.*, view-tuned units). Then, because of the maximum interaction at the afferent level, the activation pattern to the clutter ($\mathbf{A}+\mathbf{B}$) condition would be somewhere between the condition $\mathbf{A}$ and condition $\mathbf{B}$, thereby making the response of the *simple* cell to be more or less an average. The conditions for this effect are (a) some invariance operation (not necessarily a maximum) that creates an intermediate pattern of activation to the clutter condition, and (b) some tuning operation (not necessarily centered on one of the stimuli) that creates a gradient of response in the high dimensional, afferent space.

Furthermore, when there are more stimuli (> 2) presented in the clutter condition, the resulting pattern of afferent activation due to maximum-like interaction will be farther away from the optimal activation. Suppose a stimulus activates a set of afferents and compare this situation to when other stimuli are presented within its receptive field. The more cluttering stimuli there are, the more likely the afferent neuron's response will deviate from the preferred activation pattern, due to maximum-like interaction. Therefore, the response to the clutter condition with many objects will be even smaller. In other words, the tendency toward the average effect will continue to hold for not just two, but many stimuli, as shown in Fig. 4.12b.

Connection between the two complementary explanations The two mechanisms illustrated by Eq. 8 and Eq. 9 (*average by normalization*) and Fig. 4.12a (*average by invariance and tuning operations*) are closely related within the framework of our theory. As pointed out in Section 2, the invariance and selectivity operations in the model are realized by a divisive normalization circuit. The two mechanisms for producing an average response are directly implied by the architecture of the model, shown schematically in Fig. 4.13. The mechanism of normalization operation with non-overlapping set of afferents (assumptions introduced in Eq. 8) corresponds to Fig. 4.13c, and the case of max-like interaction plus tuning corresponds to Fig. 4.13a. There can certainly be intermediate cases, corresponding to Fig. 4.13b.

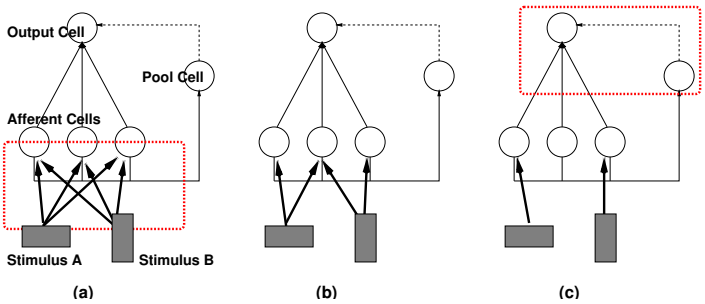


Figure 4.13: This figure shows how the two seemingly different mechanisms for the average effect are in fact closely related and both required by the theory. In each figure, a simplified model architecture is shown, which is composed of the output neuron, its afferents, and the normalizing pool cell (this is not a unique configuration, and different types of normalization circuits are also possible). In the case of Fig. 4.12a, two stimuli activate an overlapping set of afferent neurons, whose responses are determined by maximum-like operation, and the response of the output neuron is determined by the inputs from the afferents as well as the normalization, as shown in (a). In the case of Eq. 8 and 9, which corresponds to (c) and is quite similar as the proposed neural circuit for the biased competition model [Reynolds et al., 1999], the stimuli activates non-overlapping sets of afferents, and the normalization operation creates the average-like behavior. The red box indicates the main source of the average behavior in each case. (b) shows the intermediate case where two different stimuli activate two partially overlapping populations of afferent neurons.

Discussion The model, based on alternating invariance and selectivity operations, is able to account for the multiple object effect of the IT neurons in [Zoccolan et al., 2005]. However, according to the theory, not all neurons show the "average" behavior, and only under certain conditions, such effect will be observable. In general, the response of a model unit to the clutter is determined by the two key operations which are repeated throughout the model architecture. Hence, we conjecture that the multiple object effect (sometime resulting in a weighted average) in different areas of visual cortex (*e.g.,* [Reynolds et al., 1999]) may arise from the same mechanisms (see Section 4.2 on the modeling of two bar effects in V4).

A related question is how to realize clutter-invariant object recognition. Does the visual system require clutter-invariant neurons ("non-average neurons"), in order to recognize an object in a cluttered scene, or is it still possible read out the identity or category of an object from a population of clutter-intolerant neurons? A preliminary analysis (using a technique described in the next section) suggests that a population of model units, many of which are clutter-intolerant, contains enough information to perform object recognition successfully.

Notes

[15]The authors would like to thank D. Zoccolan for many fruitful discussions on this topic.

[16]The physiological recordings were performed on IT neurons showing selectivity over a pre-defined and pre-trained set of images.

[17]Counting or knowing the number of object in a cluttered scene is not trivial. Even defining "objectness" is not so simple: a face may be regarded as a single object, but, then, different parts (eyes, nose, lips, *etc.*) may be regarded as objects on their own, depending on the context. Many years of computer vision research have shown the difficulty of segmenting an object embedded in a natural background. Furthermore, our theory and the human-level performance of the model, which do not have any explicit segmentation process, show that rapid object recognition does not require such computations.

[18]Suppose that stimuli $\mathbf{A}$ and $\mathbf{B}$ activate independent, non-overlapping populations of afferents (*i.e.*, $\{x_i(\mathbf{A}) > 0\} \cap \{x_i(\mathbf{B}) > 0\} \simeq \emptyset$), so that $\sum_i x_i(\mathbf{A}+\mathbf{B}) \simeq \sum_i x_i(\mathbf{A}) + \sum_i x_i(\mathbf{B})$, which we use in Eq. 8. Such a situation would arise if the afferent neurons were rather sharply tuned to the features contained in one stimulus only. Also, we further assume that total magnitude of activation due to each stimulus alone is similar (*i.e.*, $\sum_i x_i(\mathbf{A})^q \simeq \sum_i x_i(\mathbf{B})^q$), which is a feasible situation if both stimuli are of similar structure as in the experiment. This second assumption was used in Eq. 9. Using those two assumptions, plus $r \simeq 1$, we conclude that response from a normalization operation to two stimuli would be close to an average.

4.3.3 Read-out of object information from IT neurons and from model units

Comparison of decoding performance from IT neurons vs. model units As emphasized earlier (see Section 1), one of the remarkable abilities of our visual system is the possibility of recognizing (e.g. categorizing or identifying) objects under many different views and transformations. Here we directly quantified the ability of populations of units from different stages of the model to decode information about complex objects and compared the results against recordings in IT using the same stimuli[19].

We recently used a biologically plausible statistical classifier to quantitatively characterize the information represented by a population of IT neurons about arbitrary complex objects in macaque monkeys [Hung et al., 2005a] [20]. We observed that we could accurately (performance $> 90\%$) read out information about object category and identity in very brief time intervals (as small as 12.5 ms) using a small neuronal ensemble (approximately on the order of 100 neurons). The performance of the linear classifier that we used for decoding could, at least in principle, correspond to the information available for read-out by targets of IT, such as a neuron in PFC [Miller, 2000], as discussed in Section 4.4 and Appendix A.10. During such short 12.5 ms intervals, neurons typically conveyed only one or a few spikes, suggesting the possibility of a binary representation of object features. Importantly, the population response generalized across object positions and scales. This scale and position invariance was evident even for novel objects that the animal never observed before (see also [Logothetis et al., 1995] and discussion in Section 1 and Appendix A.9).

Here we show that the observations obtained upon recording from populations of IT neurons are in agreement with the predictions made by the current theory and that there is a quantitative agreement between the model and the observations obtained from the recordings in IT. As emphasized several times throughout this report, one of the features of the model is its ability to quantitatively explore novel questions where experiments may be difficult or time-consuming. The model can thus provide quantitative predictions that can in turn help focus experimental efforts on relevant questions. After showing that the model can account for the experimental observations in IT cortex we use the model to explore several new scenarios including invariance to background changes for object recognition, invariance to presence of multiple objects and extrapolation to large numbers of objects and categories. We show that the model can account for object recognition under natural situations involving different backgrounds and object clutter and that decoding is not limited to small stimulus sets. These results are directly related to the observations discussed in Section 3 where object categorization within natural scenes was shown to work at the human performance level (in the context of an animal/not animal categorization task). We also show that the observations from IT are well accounted by the later stages of the model but not by earlier stages (S1 through S2). As expected, the layers from S1 through S2 show a weaker degree of scale and position invariance (Appendix A.9).

In order to quantitatively read out the information about objects from a population of units from the model we studied the responses from all layers of the model using the same set of 77 objects divided into 8 possible categories as used in the electrophysiology experiments described in [Hung et al., 2005a]. In order to study the degree of extrapolation to large number of objects and object categories we also used a larger set consisting of 787 images (see discussion below). Unless stated otherwise, objects were presented at the center of gaze at a size of 3.3 degrees. In order to study scale invariance we also presented the objects at 0.5x and 2x the normal scale. In order to study position invariance objects were also presented 2 degrees and 4 degrees away from the center of gaze. When studying the responses to multiple objects, objects were randomly positioned in the image with the only constraint that the objects were not allowed to overlap[21].

For all the layers, we selected units that showed activity above the response to a gray background to at least one of the 77 objects. We then randomly selected a random population of units for decoding[22]. Briefly, the input to the classifier consisted of the responses of the model units and the labels correspond to the object categories (or object identities for the identification task, see below). Data from multiple units were concatenated assuming independence. Whether this is a good assumption or not for the brain remains to be determined. Recent evidence [Aggelopoulos et al., 2005] suggests that neuronal interactions play only a small role in information encoding in IT[23]. The results shown here and the experimental results from recordings in IT suggest that we can achieve accurate read-out under this assumption. The data was divided into a training set and a test set. We used a one-versus-all approach, training 8 binary classifiers for each category against the rest of the categories and then taking the classifier prediction to be the maximum among the 8 classifiers (in the case of identification, we used 77 binary classifiers). For further details about the read-out approach see [Hung et al., 2005a] and Appendix A.9.

In Figure 4.14, we show the classification performance for the categorization task described above as a function of the number of randomly selected C2 units used to train the classifier. In agreement with the experimental data from IT, units from the C2 stage of the model yielded very high performance levels (performance $> 70\%$ for 100 units). VTU units also showed performance levels similar to those of C2 units. The performance shown in Figure 4.14 indicates the degree of extrapolation to scaled and shifted versions of the same objects. We also observed that the population response could extrapolate to novel objects within the same categories by training the classifier on the responses to 70% of the objects and testing its performance of the remaining 30% of the objects (Appendix A.9) [24].

The performance of earlier stages was significantly lower (Figure 4.14). In particular, the performance of S1 units was only weakly above chance and the performance of S2 units (the input to C2 units) was qualitatively close to the performance of local field potentials (LFPs) in inferior temporal cortex [Kreiman et al., In press][25].

The pattern of errors made by the classifier indicates that some groups were easier to discriminate than others. This was also evident from examining the correlation of the population responses for all pairs of pictures. This showed that the units yielded similar responses to similar stimuli. The performance of the classifier for categorization dropped significantly upon arbitrarily defining the categories as random groups of pictures (see AppendixA.9).

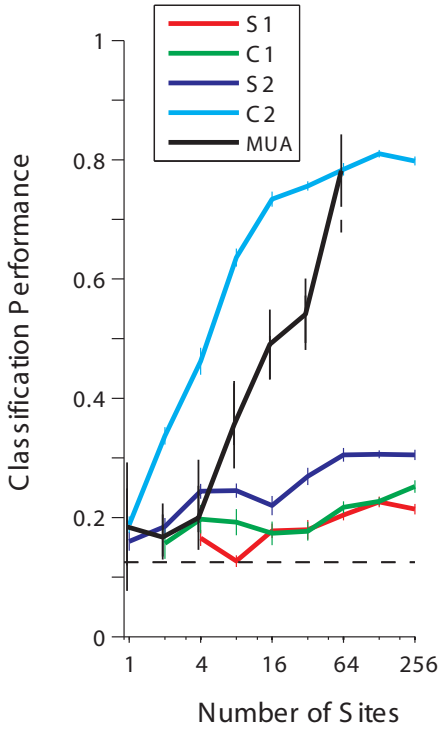


Figure 4.14: Read-out performance of a linear classifier over the entire object set on test data (not used for training) as a function of the number of units used for reading out object category for different stages of the model. Units from different model layers are shown in different colors and the black line correspond to spiking (multi-unit) data recorded from macaque IT data [Hung et al., 2005a]. The MUA data only goes up to 64 sites because we did not record IT data for 128 units in the experiment studying extrapolation scale and position invariance. The chance performance level was 12.5 % (1 of 8 possible categories, dashed line) and error bars show SD for 20 random choices of the units used for training. The 8 groups used for categorization (toys, foods, human faces, monkey faces, hand / body parts, vehicles, box outlines, cats/dogs) were defined before the experiments. The training set consisted of the 77 objects at a standard scale and position and the testing set included the same objects at two different scales and two different positions.

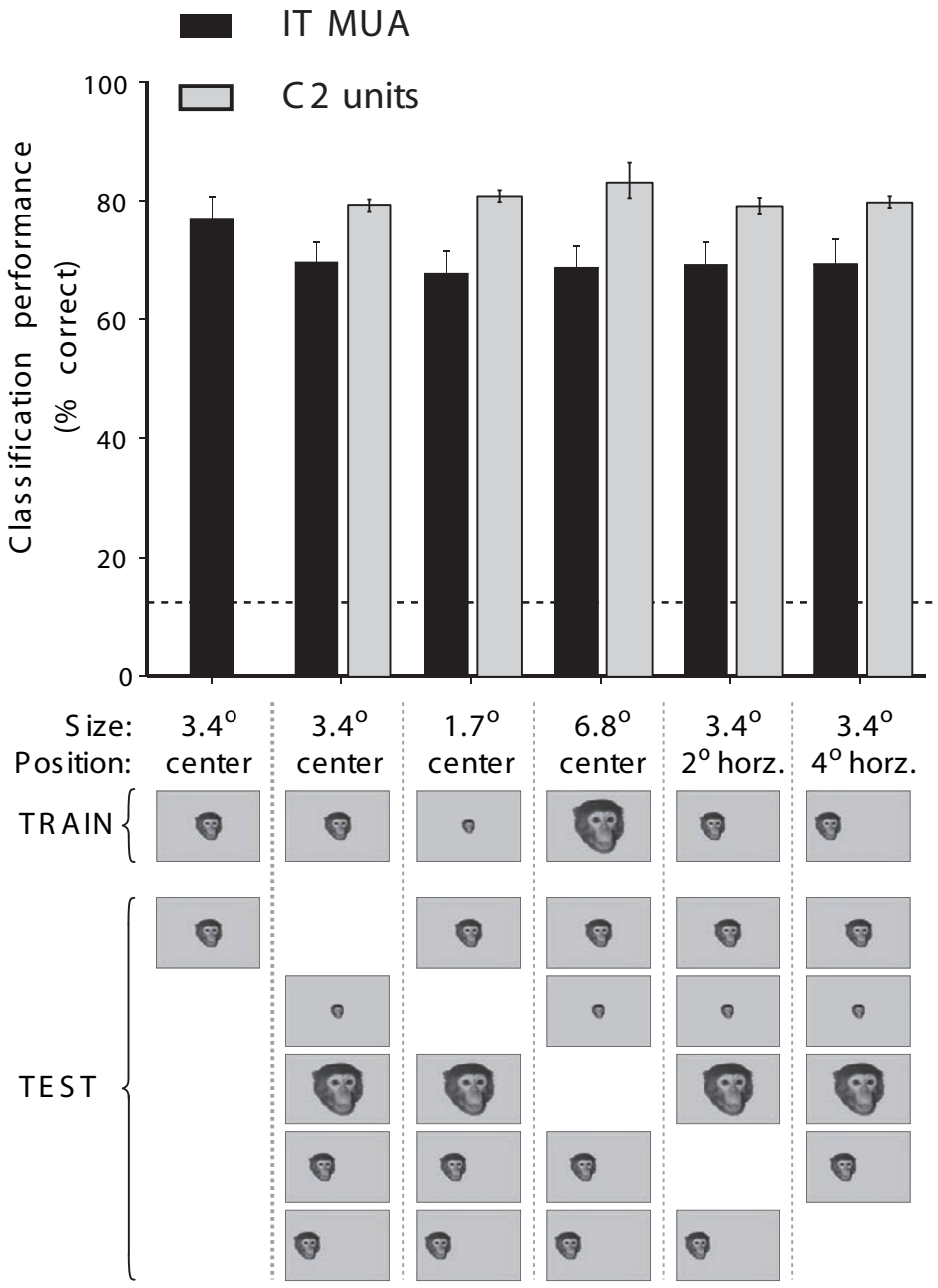


Figure 4.15: Classification performance for spiking activity from IT (black) and C2 units from the model (gray) when the classifier was trained on the responses to the 77 objects at a single scale and position (depicted for one object by 'TRAIN') and performance was evaluated with spatially shifted or scaled versions of those objects (depicted for one object by 'TEST'). The classifier never "saw" the shifted/scaled versions during training. The left-most column shows the performance for training and testing on separate repetitions of the objects at the same standard position and scale (in contrast to the model which is deterministic, neurons in IT show strong trial-to-trial variability; this bar therefore indicates the degree of extrapolation within this variability). The second bar shows the performance after training on the standard position and scale (scale = 3.4 degrees, center of gaze) and testing on the shifted and scaled images of the 77 objects. Subsequent columns use different image scales and positions for training and testing.

The performance values shown in Figure 4.14 are based on the responses of model units to single stimulus presentations that were not included in the classifier training and correspond to the results obtained using a linear classifier. Therefore, the recognition performance provides a lower bound to what a real downstream unit (e.g. in PFC, see next Section) could, in theory, perform on a single trial given input consisting of a few spikes from the neurons in IT cortex. Overall, we observed that the population of C2 units yields a read-out performance level that is very similar to the one observed from a population of IT neurons.

Robustness to stimulus transformations Many biological sources of noise could affect the encoding of information. Among the most drastic sources of noise are synaptic failures and neuronal death. To model this, we considered the performance of the classifier after randomly deleting a substantial fraction of the units during testing. As shown for the experimental data in [Hung et al., 2005a], the classifier performance was very robust to these sources of noise (Appendix A.9).

As discussed in Section 1, one of the main achievements of visual cortex is the balance of invariance *and* selectivity. Two particularly important invariances are the robustness to changes in scale and position of the images. In order to analyze the degree of invariance to scale and position changes, we studied the responses of units at different stages of the model to scaled ($0.5\times$ and $2\times$) and translated (2 degrees and 4 degrees) versions of the images. Figure 4.15 shows that the earlier stages of the model show a poor read-out performance under these transformations but the performance of the C2 stage is quite robust to these transformations as in the experimental data of Hung et al in IT [Hung et al., 2005a].

In the model, by construction, these invariances are also present for novel images. Whether a particular form of invariance is present for novel objects or not, is diagnostic of the type of learning that may be required for novel instances or transformations of those objects. We verified that IT neurons also show scale and position invariance to novel objects never seen by the monkey before the recordings ([Hung et al., 2005a], see also [Logothetis et al., 1995] and Appendix A.9).

The results shown above correspond to randomly selecting a given number of units to train and test the performance of the classifier. The brain could be wired in a very specific manner so that only neurons highly specialized for a given task project to the neurons involved in decoding the information for that task. Pre-selecting the units (e.g. using those yielding the highest signal-to-noise ratio) yields similar results but using a significantly smaller number of units[26].

In the IT recordings [Hung et al., 2005a], we devoted some effort to try to characterize the dynamics of the IT responses. In particular, we asked questions about the latency of the population response, the amount of integration time required to decode the neuronal responses and whether we could accurately detect the time of onset of the visual stimulus. Currently, the model implementation does not have any dynamics (we plan to simulate the dynamics using the circuits described in section 5).

Extrapolation to large object sets Object recognition performance for a small number of objects could potentially lead to very good results due to overtraining with specific objects or to the simplicity of the task involved. We therefore explored the performance of the model in reading out object category in a set consisting of 787 objects divided into 20 categories (the results above were based on 77 objects divided into 8 categories).

The population of C2 units conveyed information that could be decoded to identify the object's category across novel objects and locations within known categories (Figure 4.16). The classifier was trained with objects from 20 possible categories presented at different random locations and the test set included novel objects from the same categories. These results show that a relatively small neuronal population can in principle support object recognition over large object sets. Similar results were obtained in analogous computer vision experiments upon using an even larger set known as the Caltech 101 object set [Serre et al., 2005c] where an earlier version of the model could perform object categorization in a set with 101 categories. [27]

Object recognition for objects under different backgrounds The decoding experiments described above, and a large fraction of the studies reported in the literature, involve the use of well-delimited single objects on a uniform background. This is quite remote from natural vision where we typically encounter multiple objects embedded in different backgrounds, with potential occlusions, changes in illumination, etc.

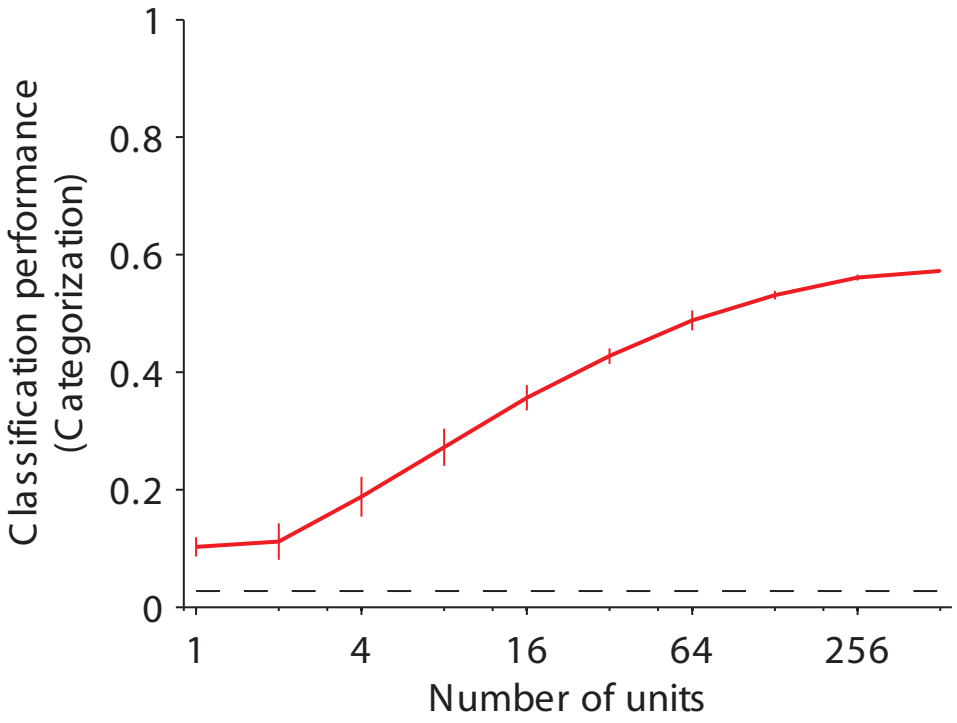


Figure 4.16: Classification performance (Categorization) as a function of the number of C2 units used to train the classifier. Here we used 787 pictures divided into 20 possible categories (cf. Figure4.14). The classifier was trained on the responses to 10% of the objects from each category (at different positions) and tested on the remaining objects. Thus, the performance results here include robustness to different objects from the same categories as well as robustness to position changes. The horizontal dashed line indicates the level of chance performance obtained by randomly shuffling the training set labels.

Ultimately, we would like to be able to read out information from IT under natural vision scenario in which an every-day life image can be presented and we can extract from the unit population activity the same type and quality of information that a human observer can (in a flash). Here we present some observations concerning the robustness of the decoding approach when objects are embedded in complex backgrounds (see also Section 3 describing the performance of the model in an animal/non-animal categorization task using objects embedded in complex backgrounds).

We presented the same 77 objects overlayed on top of images containing complex background scenes (Figure 4.17 **A**). We did not attempt to make the resulting images realistic or meaningful in any way nor did we remove the gray background surrounding the images. Particularly, when evaluating this question in humans or monkeys, the images would probably need to be appropriately modified. We used 4 different relative sizes of object to background ranging from 6 % to 69 %. The latter condition is very similar to the single object situation analyzed above both perceptually as well as from the classifier performance. The smaller relative size makes it difficult to detect the object at least in some cases when it is not salient[28]. The classifier was trained on all objects using 20 % of the background scenes and performance was evaluated using the same objects on the remaining background scenes (we used a total of 98 complex background scenes with photographs of outdoor scenes). The population of C2 units allowed us to perform both object recognition (Figure 4.17 **B**) and identification (Figure4.17 **C**) significantly above chance. Performance depended quite strongly on the relative image size (Figure 4.17 **D,E**). The largest size (69 %) yielded results that were indistinguishable from the single isolated object results shown above (cf. Figure 4.14). The small relative image size (6 %) yielded comparatively lower results but the performance of C2 units was still significantly above chance levels both for categorization as well as identification [29].

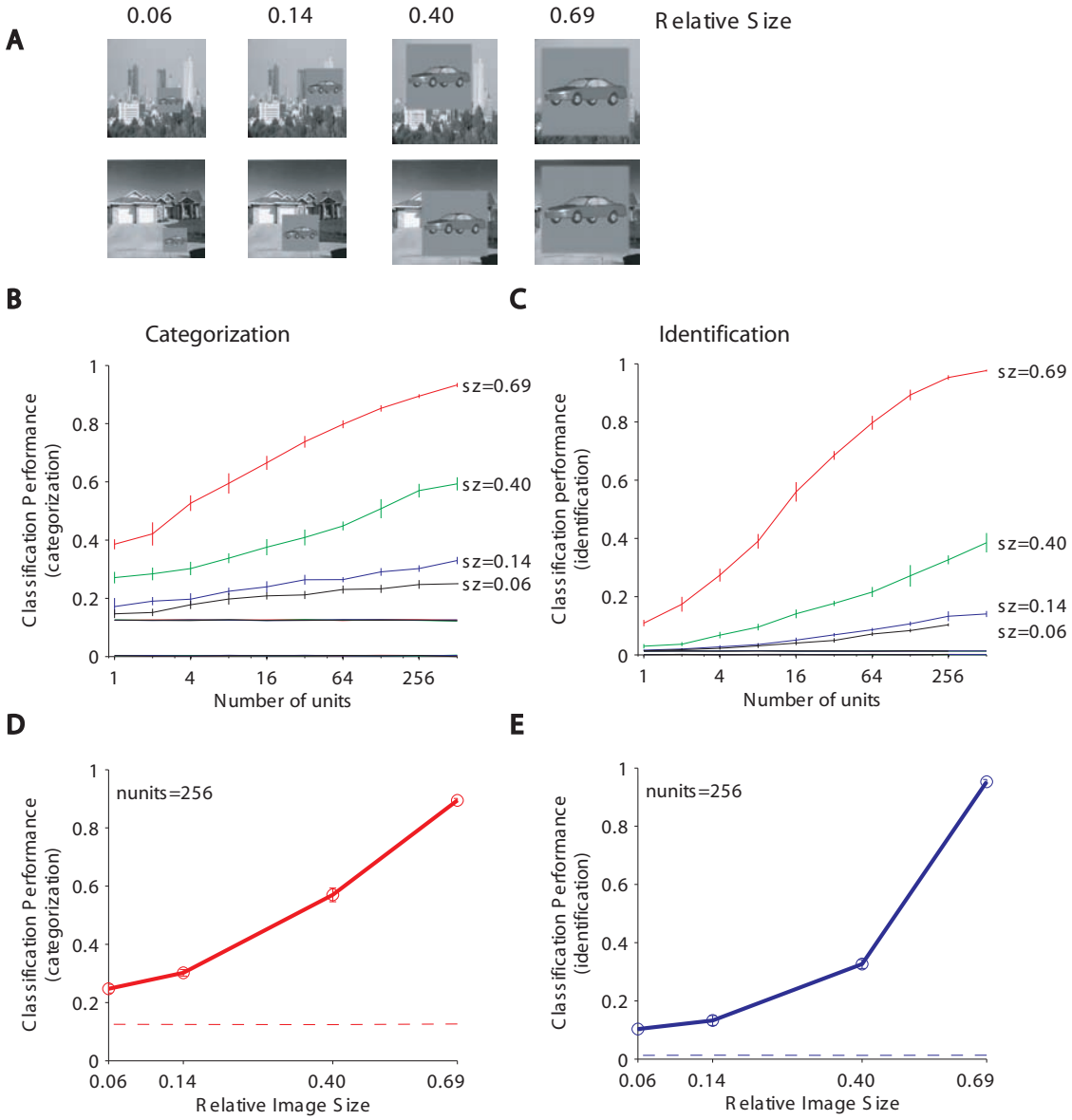


Figure 4.17: **A**. Sample of the objects pasted in complex backgrounds. Here we show a single object (a car) out of the 77 objects that were used in this experiment. Here we show the object overlayed onto two different complex background scenes (city landscape, top and house exterior, bottom) out of the 98 different background scenes that we used in this experiment. We did not attempt to generate a "meaningful" image, objects (including their surrounding gray background) were merely overlayed onto the background scenes. We used four different relative sizes of the object and background images. The center of each object was randomly positioned in the image. **B, C**. Classification performance (**B**. categorization, **C**. identification) as a function of the number of C2 units used to train the classifier. The classifier was trained using 20 % of the 98 backgrounds and the performance was tested with the same objects presented under different backgrounds. Object position within the image was randomized (both for the training and testing images). The different colors correspond to different relative sizes for the object with respect to the background. **D, E**. Classification performance (**D**. categorization, **E**. identification) using 256 units as a function of the relative size of object to background. The horizontal dashed lines indicate chance performance obtained by randomly shuffling the object labels during training.

Read-out of object category and identity in images containing multiple objects In Section 4.3.2 we described the properties of the responses of individual IT neurons and units from the model to presentation of two objects. Here we briefly present some observations concerning the robustness of the decoding approach to the presence of multiple objects.

In order to further explore the ability to decode information about an object's identity and category in natural scenes, we studied the ability to read out information from the model units upon presentation of more than one object.

We presented either two objects or three objects simultaneously in each image (Figure 4.18 **A**). We performed training either with individual objects or with images containing multiple objects (see below). During testing, the classifier was presented with images containing multiple objects. We asked several questions ranging from what is the most likely object in the image to a full description of all objects in the image (see below).

Interestingly, we could also train the classifier using images containing multiple objects. In this case, for each image, the label was the identity (or category) of one of the objects (randomly chosen so that the overall training set had the same number of examples for each of the objects or object categories). This is arguably a more natural situation under which we learn about objects since we rarely see isolated objects [30]. The performance of the classifier was only slightly higher in the single object training condition with respect to the multiple-object training condition (this depended on the number of units and the number of training examples; the largest difference observed was a performance difference of 12 %). In the figures below, we show the results obtained upon training with multiple objects.

We first considered the best prediction of the classifier (as we did above for Figure 4.14). We defined a hit in the output of the classifier if the classifier correctly predicted *either* one of the two objects presented during testing (Figure 4.18 **C**). We observed that the C2 model units yielded very high performance reaching more than 90 % both for categorization and identification with the single object training and reaching more than 80 % with the multiple object training. Given that in each trial there are basically two possibilities to get a hit, the chance levels are higher than the ones reported in Figure 4.14. However, it is clear that the performance of the C2 population response is significantly above chance indicating that accurate object information can be read-out even in the presence of another object. We also tested performance in the presence of 3 objects, obtaining similarly high results (Figure 4.18 **D**). Figures 4.18 **E-F** show the high average performance of each of the binary classifiers (8 binary classifiers for categorization (red) and 77 binary classifiers for identification (blue). The variation in performance across the different binary classifiers (i.e. across objects or across object categories) is shown in Appendix A.9.

Ultimately, we would like to be able to describe an image in its entirety. Therefore, we asked a more difficult question by requiring the classifier to correctly predict all the objects (or all the object categories) present in the image [31].

For this purpose, we took the two most likely objects (or object categories) given by the two best classifier predictions. A hit from the classifier output was defined as a perfect match between these predictions and the two objects present in the image. The higher difficulty of the task (compared to the previous tasks) is evident from the significantly lower chance levels (dashed lines in Figure A.24B,D). The performance of the classifier was also much smaller than the one reported for the single-object predictions. However, the performance was significantly above chance, reaching almost 40% for categorization (chance $= 0.0357$) and almost 8% for identification (chance $= 3.4 \times 10^{-4}$). Similar results were obtained upon reading out the category or identity of all objects present in the image in the case of 3-object images (Figure A.25).

In summary, these observations suggest that it is possible to recognize objects from the activity of small populations of IT-like model units under natural situations involving complex backgrounds and several objects. It will be interesting to empirically test these predictions.

Notes

[19] The model responses described in this sub-section correspond to the version of the model described in [Serre et al., 2005c]. It is important to note that we used that version of the model without introducing any modifications or changing any variables or parameters. Thus, the results shown here do not require the adjustment of any free parameters. The classification does require parameter adjustment but this step is analogous for the IT recordings and the model units.

[20]The electrophysiological recordings in macaque IT cortex were performed by Dr. Chou Hung in Prof. DiCarlo's lab at MIT.

[21]Occlusion is a very important phenomenon in vision under natural conditions. In many instances, parts of objects are partially occluded from view. Object recognition under normal situations can generally work in the presence of occlusion but this may depend on the quantitative details. For example, recognition under occlusion may at least partially rely on knowledge about the background in the scene. Here we did not explore the robustness of our decoding scheme to object occlusion; all objects could be seen in their entirety and there was no overlap or occlusion.

[22]Using a feature-selection method to choose the best neurons for each task based on the training data resulted in a significantly smaller required number of units to achieve the same performance, see [Hung et al., 2005a]

[23]This small contribution of neuronal interaction corresponds to analyzing the coincident spikes between neurons that are relatively far apart. It is conceivable that the interactions may depend on the distance between neurons. However, it remains challenging to simultaneously record from multiple nearby neurons and, consequently, it remains an open question whether or not the interactions among nearby neurons contribute significantly to the encoding of object information

[24]There is considerable debate regarding the nature of "categorization" and "identification" processes in the brain. The statement in the text about extrapolation to novel objects within the same category refers to the possibility of decoding the presence of a similar object which was *not* presented during training but belonging to the same categories used during training (e.g. training on several human faces and then categorizing a novel face never seen during training). This should not be taken to imply that IT neurons (or C2/VTU units) are "category" neurons in the sense that their responses depend on the category assignment regardless of identity consideration, or that there are specific brain regions devoted to these categories. From a computational viewpoint, categorization and identification appear to be part of the same process and we try to emphasize throughout the text a general recognition machinery which is capable of performing all recognition tasks. In the model, the weights from VTU to PFC can be adjusted to perform one task or the other. Recordings from macaque PFC suggest that PFC neurons are sensitive to the task and that their responses represent category limits (see [Freedman et al., 2001, 2003] and Appendix A.9).

[25]The main components of local field potentials are dendritic potentials and therefore LFPs are generally considered to represent the input and local processing within a cortical area [Mitzdorf, 1985; Logothetis et al., 2001]. Thus, it is tempting to speculate that the LFPs in IT are closer to the responses of S2 units in the model. However, care should be taken in this interpretation since the LFPs constitute an aggregate measure of the activity over multiple different types of neurons and large areas. Further investigation of the nature of the LFPs and their relation with the spiking responses could help unravel the transformations that take place across cortical layers.

[26]This could be a role for selection mechanisms including attention. For example, when searching for the car keys, the weights from some neurons could be adjusted so as to increase the signal-to-noise ratio for specific tasks. This may suggest that other concomitant recognition tasks would show weaker performance. In this case, the selection mechanisms take place before recognition by biasing specific populations for certain tasks; this does not imply that feedback mechanisms are operating during the recognition process.

[27]How many objects or object categories could be encoded and decoded with this model? It would seem that humans have a very large capacity for storing different objects and capacities (the exact limits of which are hard to define). The results shown here based on the model, together with those obtained upon decoding the responses from an IT population in [Hung et al., 2005a] and the simulations in [L.F.Abbott et al., 1996] suggest that the number of objects (or classes) that can be decoded at a given accuracy level grows approximately as an exponential function of the number of units. Even allowing for a strong redundancy in the number of units coding each type of feature, these results suggest that networks of thousands of units could display a very large capacity. There are multiple assumptions in this argument but, at the very least, there does not seem to be any obvious capacity limitations for hierarchical models to encode large numbers

of objects and categories.

[28]A more rigorous quantification of this effect was discussed in the context of the comparison with human performance, see Section 3. The psychophysics results shown there showed that the "far body" condition yielded a significantly lower performance than the "close body" condition for an animal/non-animal categorization task.

[29]Recognizing (and searching for) small objects embedded in a large complex scene, e.g. searching for the keys in your house, is actually one example of a task that may require additional resources. These additional resources may involve serial attention which is likely to be dependent on feedback connections.

[30]It is possible that attentional biases to some extent "isolate" an object, e.g. when learning about an object with an instructor that points to it.

[31]We generally live under the illusion that we can recognize and describe every object in the image during a glimpse. Multiple psychophysics studies suggest that this is probably wrong. Perhaps one of the most striking demonstrations of this is the fact that sometimes we can be oblivious to large changes in the images (see for example [Simons and Rensink, 2005]). What is the capacity of the representation at-a-glance? There is no consensus answer to this question but some psychophysical studies suggest that only a handful of objects can be described in a brief glimpse of an image (on the order of 5 objects). After this first glance, eye movements and/or attentional shifts may be required to further describe an image. We continue here referring to this rapid vision scenario and we strive to explain our perceptual capabilities. Thus, the goal is to be able to fully describe a set of about 5 objects that can be simultaneously presented in multiple backgrounds in a natural scenario.

4.4 PFC

The model – in the present incarnation – assumes that task-dependent learning appears between IT and PFC and not earlier: IT is the last purely visual area which is task independent.

The experimental evidence seems to broadly support this view – which is certainly simplified. PFC cells responses are much more task dependent than responses of IT cells [Freedman et al., 2003]. Recent recordings [Freedman et al., 2001, 2002, 2003] revealed that neurons in PFC were often "category-tuned", conveying reliable information about category membership – learned in a supervised way – and relatively little information about individual stimuli within each category. By contrast, the majority of neurons in ITC showed shape-tuning; they tended to show selectivity for individual stimuli and little evidence for selectivity based on category membership per se (see Figure 4.19). Also responses in IT seem to be similar in passive vs. active viewing, whereas responses in PFC vary dramatically switching from passive to active viewing.

The present model assumes that classifier circuits are in PFC (and probably other higher visual areas). The evidence from the same papers [Freedman et al., 2001, 2002, 2003] and from model simulations [Riesenhuber and Poggio, 1999a] supports this aspect of the model since responses of a categorical type appear in PFC.

This paper does not attempt to describe how PFC works. This problem may be part of a future theory incorporating higher-level processes, backprojections and their control. As we mentioned, the theory assumes that PFC can "read out" the appropriate information from IT by setting up – for instance through supervised learning – task-dependent classifier *circuits*, in principle as simple as a linear classifier. A linear classifier – depicted in the model of Figure 2.1 as the top connections from the highest units in IT to circuits for category, identification, expression estimation and a large number of other learned tasks – could be implemented by a set of synaptic weights on a single PFC neuron and by its threshold. It is more likely that in reality more complex circuits are used. In any case recent read-out experiments have shown that several different kinds of information, – such as identity, category, position, size – can be read out *during passive viewing* from a small number of IT neurons by a linear classifier, providing a plausibility proof of our simple hypothesis [Hung et al., 2005b,a].

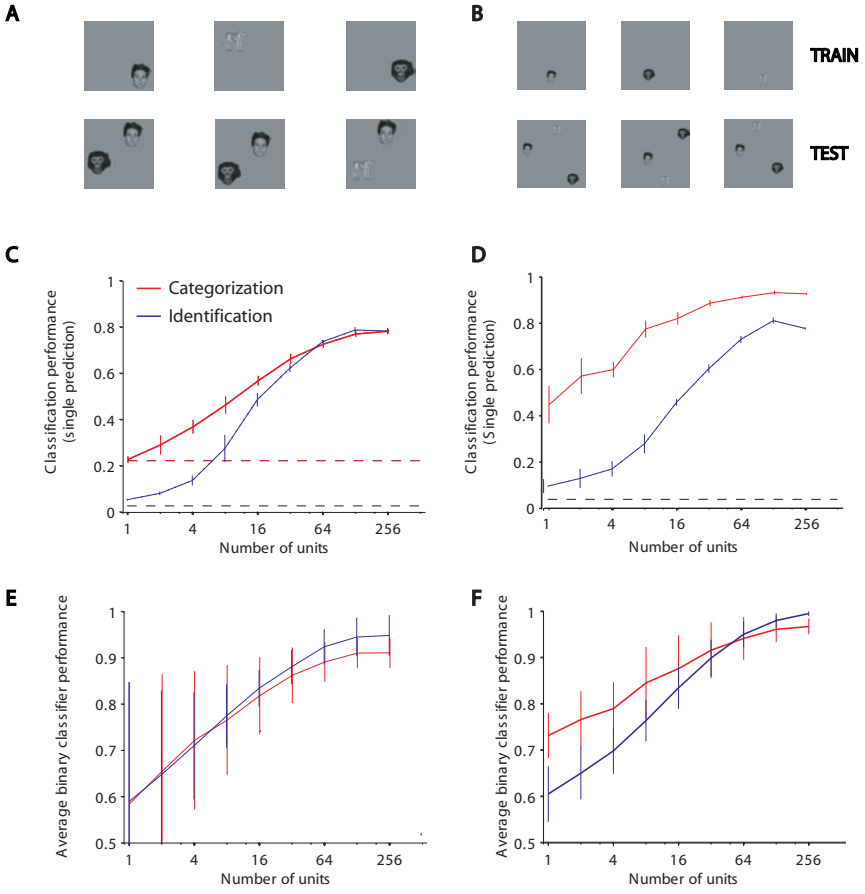


Figure 4.18: Classification performance for reading out object category (red)or object identity (blue) in the presence of two objects (**A, C, E**) or three objects (**B, D, F**). **A, B** Examples of the images used in training (top) and testing (bottom). Here, we show images containing single objects to train the classifier (top). However, performance was not significantly different when we used images containing multiple objects to train the classifier (see text and Appendix A.9 for details). **C, D** Classification performance as a function of the number of C2 units used to train the classifier. Here we used a multi-class classifier approach; the output of the classifier for each test point was a single possible category (or object identity) and a we considered the prediction to be a hit if this prediction matched *any* of the objects present in the image. The dashed lines show chance performance levels and the error bars correspond to one standard deviation from 20 random choices of which units were used to train the classifier. We exhaustively evaluated every possible object pair or triplet. **E, F** Average performance for each of the binary classifiers as a function of the number of C2 units used for training. The number of binary classifiers was 8 for categorization (red) and 77 for identification (blue). The error bars show one standard deviation over 20 random choices of C2 units.

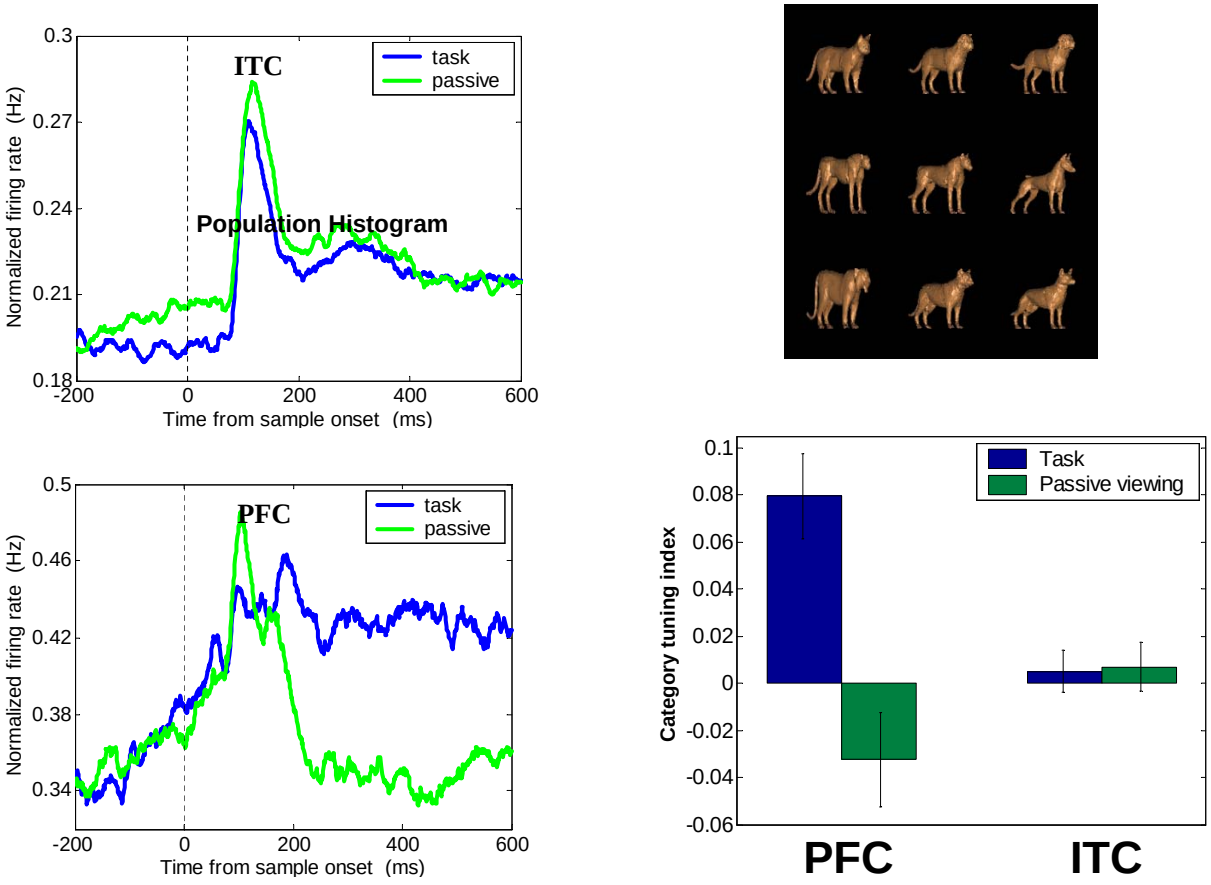


Figure 4.19: Are stimulus representations in PFC and ITC modulated by changes in task demands? Simultaneous recordings from 298 ITC, 212 PFC neurons from two monkeys were made by Freedman *et al.* (in prep.) while monkeys alternated between categorization task and passive viewing (using cats and dogs images, see Section A.10). The figure shows that ITC activity was similar between task and passive viewing while PFC responses were more task-dependent. As shown on the panel on the bottom right, PFC category tuning was stronger during task while ITC category tuning was weaker and not affected by the task.

5 Biophysics of the 2 basic operations: biologically plausible circuits for tuning and max

To support the biological plausibility of the theory it is important to show that the two basic operations required by it can be implemented by known or plausible properties of neurons and synapses. Thus the main goal of this chapter is to describe plausible circuits for the max and the tuning operations. Interestingly, it turns out that there are several possible circuits apparently working at the level of robustness required by our model simulations. Thus the circuits described here also represent specific hypotheses that neurophysiologists can try to prove or disprove using, for instance, extracellular and intracellular recordings.

Several groups have reported that some neurons in visual cortex respond rapidly and sublinearly to the combined presentation of two simple stimuli in their receptive field [Gawne and Martin, 2002; Lampl et al., 2004], i.e. the response to the combined stimulus is significantly smaller than the sum of the responses to the single stimuli. It has been proposed that, instead of summing the inputs, these neurons compute either the maximum or the average of the inputs. The computational motivation for the maximum arises from the need to preserve specificity while building up invariance in the ventral visual stream [Riesenhuber and Poggio, 1999b] while normalization circuits are required to implement tuning of S units in the model. Normalization circuits were suggested (for gain control) in [Carandini and Heeger, 1994] and (for the biased competition model) in [Chelazzi et al., 1998] (see also [Poggio et al., 1981; Reichardt et al., 1983] and for older work on related dynamics of shunting inhibition [Grossberg, 1973]). Another possible mechanism for gain control relies on synaptic depression [Abbott et al., 1997]. Several possible circuits for computing the maximum operation have been proposed on an abstract level [Yu et al., 2002; Poggio et al., 1981], but not investigated with respect to their biophysical plausibility. Based on the general architecture of those circuits, we present here examples of a few biophysically plausible models of microcircuits for computing the maximum and the normalization operation – which is assumed to be at the core of the tuning (see section 2).

One of the most important constraints for the presented models stems from the dynamics of recorded cells. Within 20ms of response onset, recordings of V1 cells for instance show differential depolarization and firing rates, respectively. Thus, the underlying circuits must be capable of responding to different inputs in a differential manner within 20ms. Based on the peak firing rates of neurons in V1 of no more than 100Hz, only one or two spikes can be transferred during that time. This poses a problem since such a binary signal is not sufficient to explain the variation in responses. There are two main possibilities that we investigate: nonspiking neurons featuring graded synaptic transmission and ensembles of equivalent spiking cells.

Another constraint from the model for the two operations is the dynamic range which is required for performance of the model, in other words the precision with which tuning and max have to be performed. We know that the model is robust to perturbations of the operations and in particular we know that the output at the level of C3 can be binarized without destroying the performance of the model in recognition tasks. We did not perform an estimate of required dynamic range at each level of the model, though we plan to do it soon.

All the models described here follow the idea of an interplay of balanced excitation and inhibition in cortical microcircuits and even single cells [Mariño et al., 2005; Chance et al., 2002; Douglas et al., 1995; Hahnloser et al., 2000] and emphasize the importance of the synapse as computational entity, in the spirit of [Abbott and Regehr, 2004].

We will start describing non-spiking circuits for the normalization and the maximum operation, followed by spiking circuits. Needless to say, most cortical physiologists believe that most neurons in cortex spike, though we prefer to keep a healthy skepticism for now, since neurons with graded synaptic transmission in cortex may have not been found due to the bias to look for spiking cells and graded synaptic transmission has been found and characterized in invertebrates [Manor et al., 1997]. A more detailed quantitative description of the models can be found in appendix A.11.

5.1 Non-spiking circuits

For all the circuits described here, we use a single compartment model of a cell with its membrane potential as the main variable. Thus there are no delays or other cable effects for dendrites or axons at this point. Synapses are assumed to directly connect the somata of the two involved cells. Each synapse is modeled

in two stages: the transmitter release and the receptor binding. First the presynaptic potential is converted to a neurotransmitter concentration. In a second step, the neurotransmitter binding to the receptors results in a conductance change. For the nonspiking circuits, we assume that the release of neurotransmitter at the synaptic terminal is a graded and continuous function of the graded depolarization of the cell.

As input to the circuit, we set the membrane potential of the x units at different values. Based on that, the transmitter concentration and postsynaptic conductance for all synapses and the potential for the p, y and z cells are computed.

5.1.1 Normalization

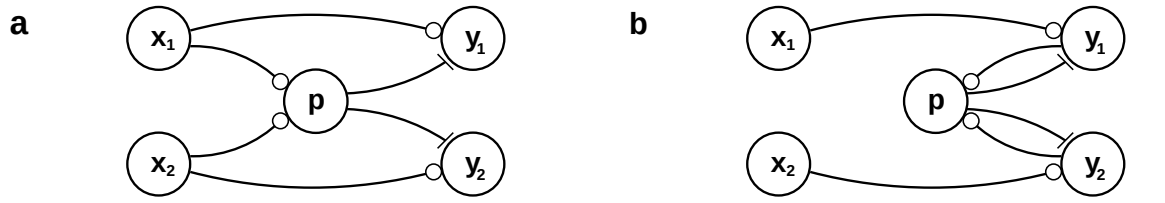


Figure 5.1: Network architecture for normalization. a) Feedforward. b) Feedback. Small circles indicate excitatory synapses, bars represent inhibitory connections.

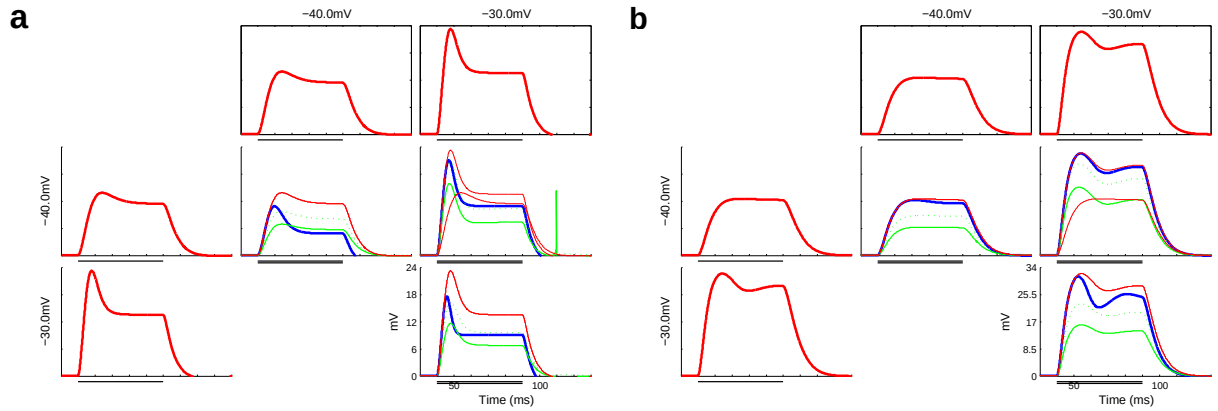


Figure 5.2: Membrane potential of a y unit. Feedforward model (left panel) and feedback model (right panel). The plotted traces show the membrane potential of the z unit over time. The strength of the inputs A and B is varied along the axes. The upper- and leftmost red traces show the response of the circuit to only stimulus A or B, respectively. The black line below the x-axis indicates the duration of the stimulus. In the other traces, the red traces correspond to the responses to only A and B as in the upper- and leftmost traces. The blue curve is the actual response to the combination of stimuli A and B. For comparison, the solid green line shows the L1 normalized value and the dotted green line shows the L2 normalized value. The two black lines underneath the traces indicate the duration of stimuli A and B.

The basic circuit is shown in Figure 5.1. Each input neuron x_i excites the corresponding neuron y_i. In addition, each x_i excites a common pooling interneuron which inhibits – thereby normalizing, as we will show – each y_i. We studied both feedforward (Fig. 5.1a) and feedback circuits (Fig. 5.1b). Variations of this model include a lateral inhibition circuit in which there is an inhibitory interneuron for each x_i inhibiting each y_j and even a version thereof with direct all-to-all lateral inhibition without interneurons.

As shown in Fig. 5.2, the y units of the feedforward nonspiking normalization circuit exhibit a response very close to the L2-normalized input while the feedback version does not show this behavior.

5.1.2 Max

The circuits for the Max operation are very similar to the circuits for normalization. They are based on the architecture presented in [Yu et al., 2002]. There is now an additional output unit z which is excited by all the y_i. We studied both feedforward (Fig. 5.3a) and feedback inhibition (Fig. 5.3b). As in the case of normalization, variations of this model include a lateral inhibition circuit in which there is an inhibitory interneuron for each x_i inhibiting all y_j and even a version thereof with direct all-to-all lateral inhibition without interneurons. We make the same assumptions and use the same equations as for the normalization circuits.

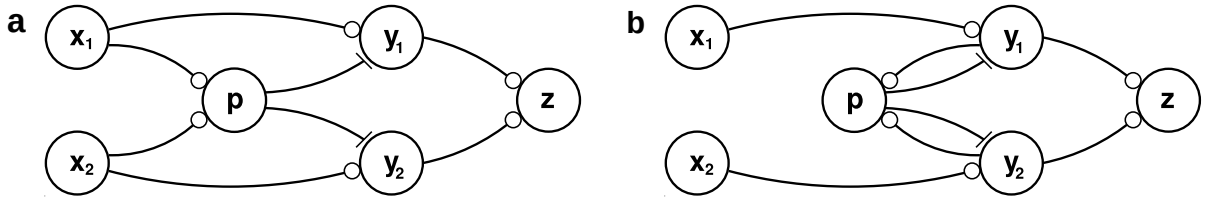


Figure 5.3: Network architecture for Max. a) Feedforward. b) Feedback. Small circles indicate excitatory synapses, bars represent inhibitory connections.

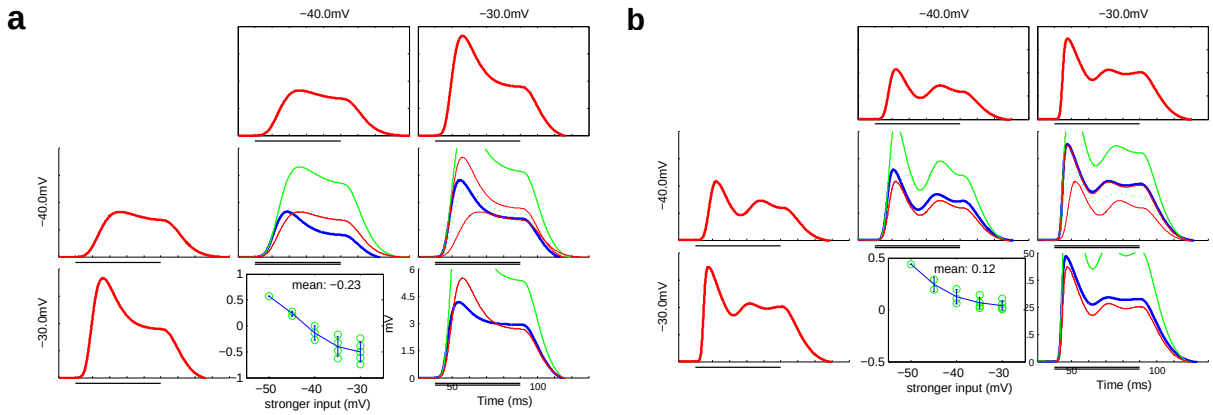


Figure 5.4: Membrane potential of a z unit. Feedforward model (left panel) and feedback model (right panel). The plotted traces show the membrane potential of the z unit over time. The strength of the inputs A and B is varied along the axes. The upper- and leftmost red traces show the response of the circuit to only stimulus A or B, respectively. The black line indicates the duration of the stimulus. In the other traces, the red traces correspond to the responses to only A and B as in the upper- and leftmost traces. The blue curve is the actual response to the combination of stimuli A and B. For comparison, the solid green line shows the sum of the two red curves and the dotted green line shows the average. The two black lines underneath the traces indicate the duration of stimuli A and B. The inset shows the Sato max index [Sato, 1989] for all the input configurations in the respective panel, grouped by the size of the stronger input: $I = \frac{R_{AB} - max(R_A, R_B)}{min(R_A, R_B)}$.

Feedforward model Figure 5.4a shows the response to two coincident stimuli with synaptic strengths of $20nS$ for both excitation and inhibition. For small inputs, the response is approximately flat. The combined response traces the sum of the single responses or is slightly sublinear. However, if both inputs are considerably large, we observe a fast transient strong response that is followed by a longer plateau because the inhibition is slightly delayed. This shape is characteristic for any delayed suppression mechanism. Synaptic depression causes a similar response pattern, but on a slower timescale [Chance et al., 1998]. The combined response to two strong inputs is even below the maximum of the individual responses, closer to the average of the individual responses. However, for most simulations, the combined response seems to be

consistently sublinear and depending on parameters it can be either closer to the max or to the average. It is also interesting to note that despite very different transient responses, the height of the plateau is almost constant for different strong inputs. If both inputs are equal, the combined response is transiently below the individual responses but the plateau is higher.

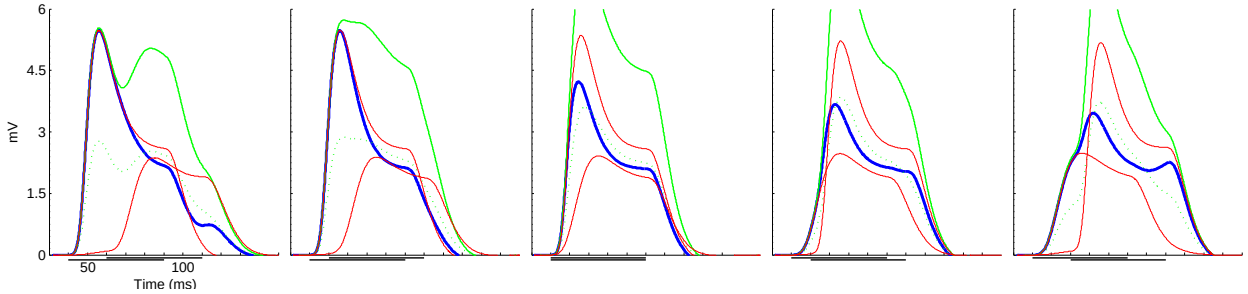


Figure 5.5: Membrane potential for different delays (-20ms, -10ms, 0ms, 10ms, 20ms) in the feedforward model.

To study the effect of asynchronous inputs, we delayed the weaker and stronger inputs by up to 20ms in figure 5.5. If the stronger input arrives first, the combined response traces the stronger individual response pretty well, except for the time when the slower response is much stronger towards the end of the simulation, where the combined response resembles the average (this is less the case for weaker inputs, as always). If the weaker input arrives first, the combined response traces the weak input as long as the stronger input has not set on. As soon as the stronger input sets on, the combined response resembles the average of the individual responses.

Feedback model Figure 5.4b shows the response to two coincident stimuli with synaptic strengths of $100nS$ for both excitation and inhibition. Small inputs elicit a somewhat symmetric hill-like response. Stronger inputs show the same strong transient/lower plateau behavior as the feedforward model. The combined response resembles the maximal individual response and is only slightly higher for two strong inputs.

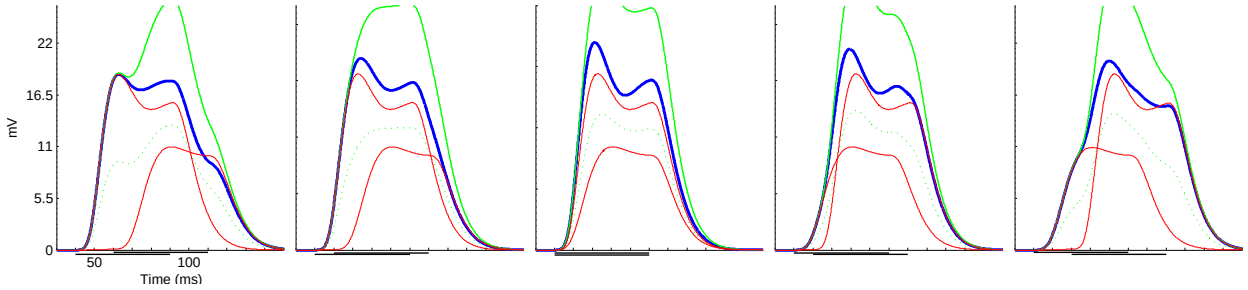


Figure 5.6: Membrane potential for different delays, feedback model, $\bar{g}_{AMPA} = 20nS$, $\bar{g}_{GABA_A} = 20nS$. Stimulus length: 50ms

As for the feedforward model, we also tested the response to delayed inputs. Unlike the feedforward circuit, this version shows very little modulation in relation to the timing of the inputs (cf. Fig. 5.6).

Architecture variations We also investigated architectural variants with lateral inhibition with and without interneurons for both feedforward and feedback models. As in the presented results, the feedback circuit is more robust and close to a maximum regardless of the exact parameter settings whereas the feedforward circuit's behavior varies more with different parameters and covers a range of sublinear regimes.

More than two inputs Given the highly convergent architecture of the brain, which is also reflected in the presented model of object recognition, it seems unlikely that cortical circuits compute the maximum over only two inputs. In order to test the plausibility of our circuits for a larger number of inputs, we extended the circuit by adding more x_i and y_i units and the corresponding synapses to include an arbitrary B number of inputs.

Preliminary simulations with 5 and 10 inputs, respectively, show that the investigated class of extended models exhibits sublinearity in the case of multiple inputs. In particular, the output can significantly deviates from the maximum for 5 and more inputs. This poses a challenge for these circuits and we are currently investigating this problem. As we mentioned earlier, this problem is closely related to the dynamic range required in the model at the level of the maximum operation.

5.2 Spiking circuits, wires and cables

Most cortical physiologists believe that most neurons in cortex spike.

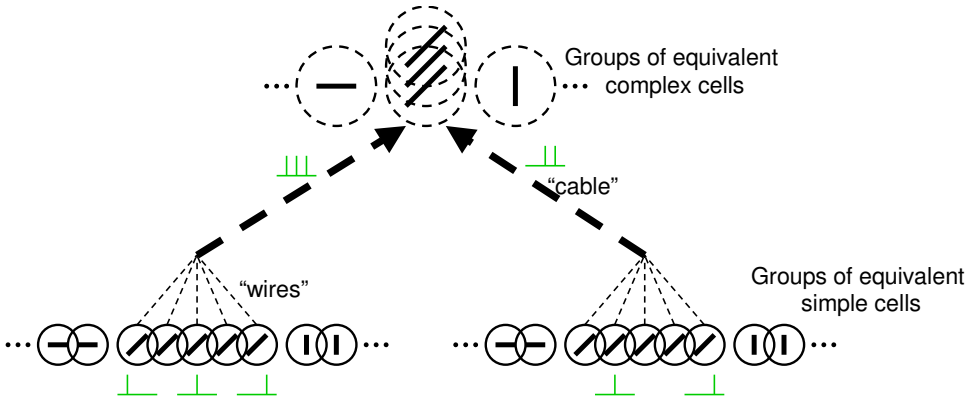


Figure 5.7: Signal propagation between two somata with spiking neurons and "cables". Each of the "equivalent" cells on the bottom, which receive identical input, spikes independently. Each spike is propagated independently in its own "wire" in the "cable" and then triggers neurotransmitter release at one synapse for each of the equivalent postsynaptic neurons. This process is used for each synapse in the model.

In the spiking circuits, a natural solution to the problem of insufficient dynamic range is to replace each unit with a group of n "equivalent" cells. All equivalent cells have identical parameters and receive identical inputs from the other units in the circuit, but in addition each cell receives individual normally distributed background input. Instead of 2 spikes, the postsynaptic cell will now receive up to $2n$ spikes, *i.e.,* the dynamic range of the input is multiplied by n, as shown in Fig. 5.7. The number of equivalent cells n probably decreases along the visual hierarchy from V1 to IT. In early stages, a large dynamic range of the inputs is needed, whereas at the other extreme in IT, only the binary presence or absence of each feature has to be conveyed.[1]

Each of the input x_i units, i.e. each of the equivalent cells, receives a square pulse current input of equal length (10ms) with normally distributed amplitude and onset.

For each arrow in figures 5.8 and 5.9 there are n equivalent leaky integrate-and-fire cells and thus n synapses at each postsynaptic cell. Each of these synapses is modeled independently, contributing to the total synaptic conductance of the cell. In addition to the inputs from other cells in the circuit, each postsynaptic cell receives normally distributed background conductance input.

5.2.1 Normalization

The architecture of the simplified spiking circuit for normalization is depicted in Fig. 5.8. We are currently investigating the properties and performance of this circuit.

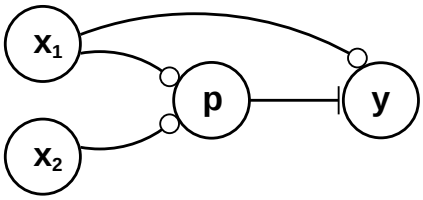


Figure 5.8: Simplified feedforward normalization network architecture. Small circles indicate excitatory synapses, bars represent inhibitory connections. We assume that units and connections follow the "cable" architecture depicted in Fig. 5.7.

5.2.2 Max

Based on the data from experiments performed in Ilan Lampl's lab [Levy and Lampl, 2005] we implemented a noisy spiking version of the model with a slightly changed architecture as depicted in Fig. 5.9. In this version, the x units directly excite the z unit and the pool unit also inhibits z directly. This is more consistent with measurements of excitatory and inhibitory input currents in rat barrel cortex (S1).

Figure 5.9: Simplified feedforward (a) and feedback (b) max network architecture. Small circles indicate excitatory synapses, bars represent inhibitory connections.

Figures 5.11 through 5.13 show the traces of the mean membrane potential of the x_i and z units as well as the excitatory xz and inhibitory pz conductances at the z unit. While one input is held constant, the other input is increased from one figure to the next. In the conductance plots, a significant change in the ratio of excitatory and inhibitory conductances is evident. This is consistent with data from [Levy and Lampl, 2005].

Figure 5.14 plots the mean membrane potential of the z unit for different combinations of input currents shown in the legend (in nA). While the peak depolarizations are almost identical, stronger combined inputs yield a faster rise time of the membrane potential. If only one of the inputs is active, the response is also about twice as broad.

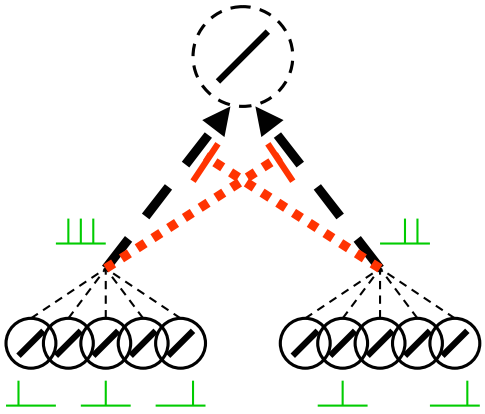


Figure 5.10: Possible spike-timing dependent max circuit. Each group of cells inhibits the other competing ones. The group with the first spike wins and blocks all other competitors.

An alternative spike-timing based circuit for max Based on observations in our own simulations and the experiments done by Tim Gawne [Gawne, 2000] that stronger inputs lead to earlier spikes (conforming

with an earlier proposal [Thorpe et al., 1996]), another possible mechanism for computing the maximum of a number of inputs is that the first and thus strongest input shunts all later inputs. There are several possible implementations of this idea. One possibility consists of inhibition at the presynaptic axon, another one employs dendritic circuits similar to "dendritic motion detectors". Depending on the implementation, there are strong constraints on the timing of inhibition, jitter between equivalent units and timing differences between the inputs. Also, for inputs of nearly identical strength, most of these implementations will most likely fail to preserve the maximum output. Simulations to confirm these conjectures are pending.

5.3 Summary of results

Overall, the nonspiking feedforward circuit is able to produce different types of responses based on its parameters. It seems to be particularly suitable for the normalization operation. The nonspiking feedback circuit is more invariant to changes in parameters such as synaptic strengths and input delays. The combined response *is always very close to the maximum of the two individual responses* at the current time for strong inputs and traces the sum for weak inputs. We do not have simulation results for the spiking normalization circuit yet, but first simulations of the spiking max circuit show promising results compatible with data from [Levy and Lampl, 2005]. An open challenge under investigation is the problem of multiple inputs.

5.4 Comparison with experimental data

Figure 5.15 shows the instantaneous firing rate over time in response to visual stimuli of different contrast. For high contrasts, a high transient response is follow by a lower sustained response. Lower contrasts show a slower increase in firing rate and don't exhibit a peak before the plateau. This behavior is most likely due to a time-delayed suppression mechanism that is coupled to the activity of the circuit. We see the response shape of the membrane potential in our nonspiking models due to the slightly delayed inhibition in the circuits (cf. Fig. 5.2 and 5.4).

In figure 5.16, two stimuli elicit a short latency and long latency response, respectively. The short latency response is a high transient response while the long latency response is lower but sustained for a longer time. The response to the combined stimulus strongly resembles the single short latency response and does not exhibit a slower sustained component. Thus, the short strong transient response blocks the slower response. In the model, we see similar effect if we artificially delay the smaller input by 20ms (cf. Figure 5.5).

Intracellular recordings by Lampl et al. are shown in Figure 5.17. These recordings show that during presentation of the combined stimulus, the membrane potential of these cells traces the response to the stronger stimulus only. This data was one of the starting points for our models which produce similar behavior for several parameter settings.

5.5 Future experiments

In order to decide which of the presented models and variations, if any, best describes the biophysical reality, new experiments are need that help to distinguish between the alternatives.

A very straightforward experiment would be to change the timing of the presented visual stimuli to test the temporal precision that is needed for the maximum effect to occur and what kind of behavior can be observed outside of that precision window.

Another interesting experiment would be to extend from two to three or even more inputs, i.e. stimuli. This is probably difficult because of the small receptive field sizes of the cells under investigation but it would help to tease apart different possible mechanisms as their behavior for more than two inputs can be very different.

Another idea might be the direct stimulation of either preserved cortical microcircuits in slices or possibly cell cultures in order to further investigate the principal ways in which different inputs to a cell are combined.

As a last point, given the current model microcircuits predicting nonspiking graded transfer of signals across chemical synapses, it might be worthwhile to take another look at cortical networks with the aim to look for neurons and synapses of this type.

Notes

[1]Contrast invariance data provide some indirect support to the idea of the *cables getting thinner along the hierarchy*. [Sclar et al., 1990] showed that the steepness of the contrast-response functions of neurons increases from LGN through V1, V2 to MT and that "cells become, in the contrast domain, progressively more like switches, being either on or off" [Lennie, 1998].

[2]We would like to thank David Ferster, Ian Finn and Ilan Lampl for providing their recorded data. Additionally, we would like to thank them as well as Christof Koch for numerous fruitful discussions.

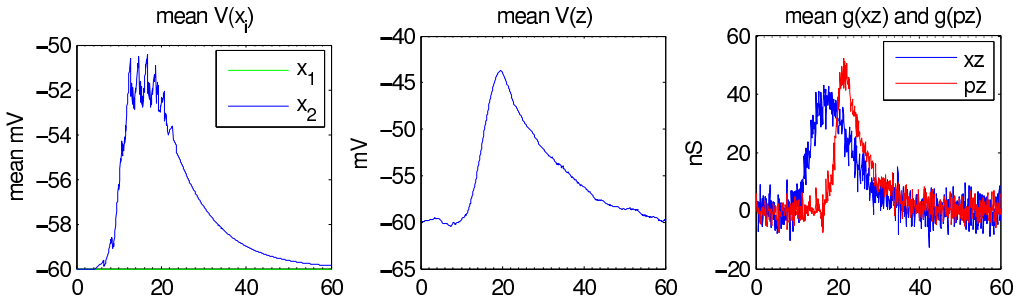


Figure 5.11: Mean x_i and z unit depolarizations and sum excitatory (xz) and inhibitory (pz) conductances at the z unit for input 1: $0nA$ and input 2:$4200nA$.

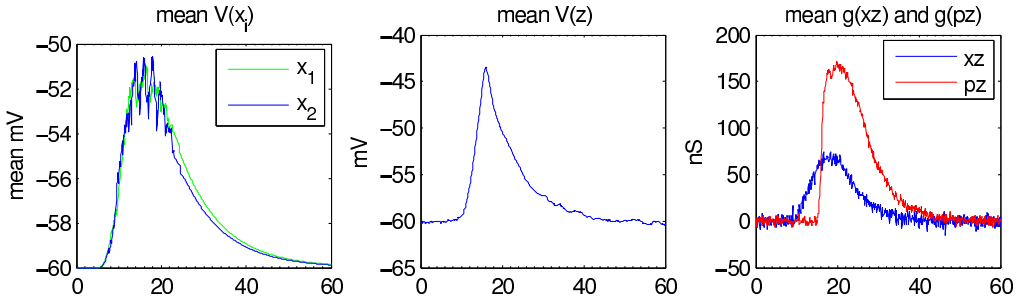


Figure 5.12: Mean x_i and z unit depolarizations and sum excitatory (xz) and inhibitory (pz) conductances at the z unit for input 1: $3000nA$ and input 2:$4200nA$.

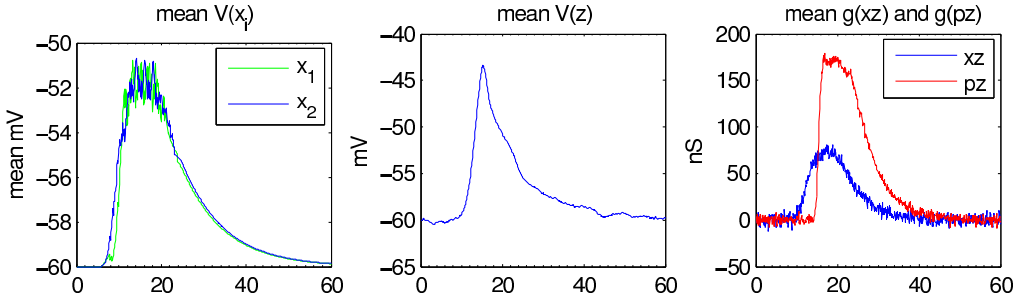


Figure 5.13: Mean x_i and z unit depolarizations and sum excitatory (xz) and inhibitory (pz) conductances at the z unit for input 1: $4200nA$ and input 2:$4200nA$.

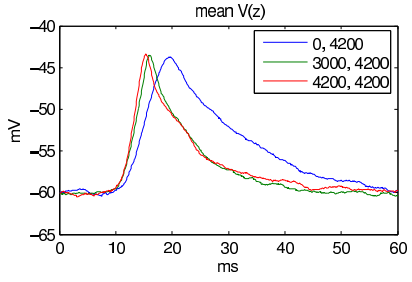


Figure 5.14: Mean z unit depolarizations for different combinations of input currents noted in the legend (in nA).

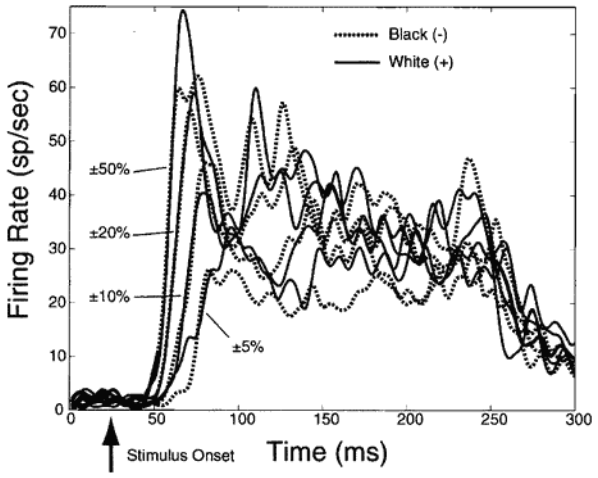


Figure 5.15: Instantaneous firing rates of complex cells in primate V1 in response to stimuli of different contrasts. [Gawne, 2000]

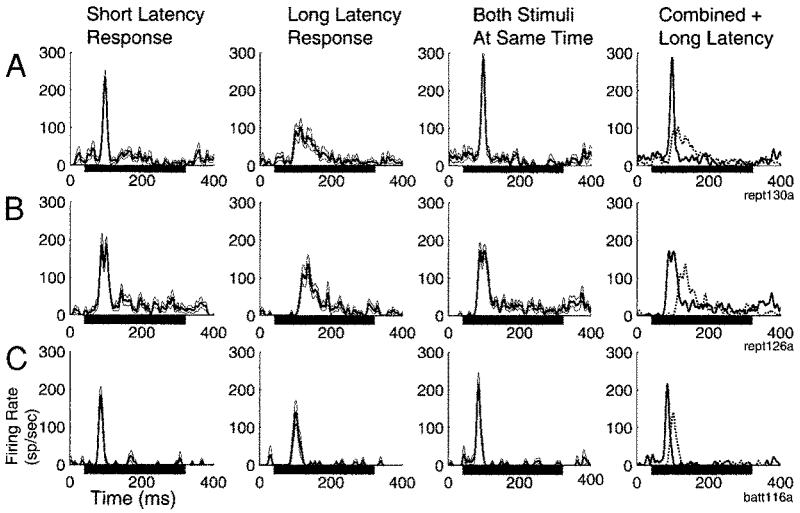


Figure 5.16: Instantaneous firing rates of a primate V4 cell in response to the combined presentation of two stimuli. [Gawne and Martin, 2002]

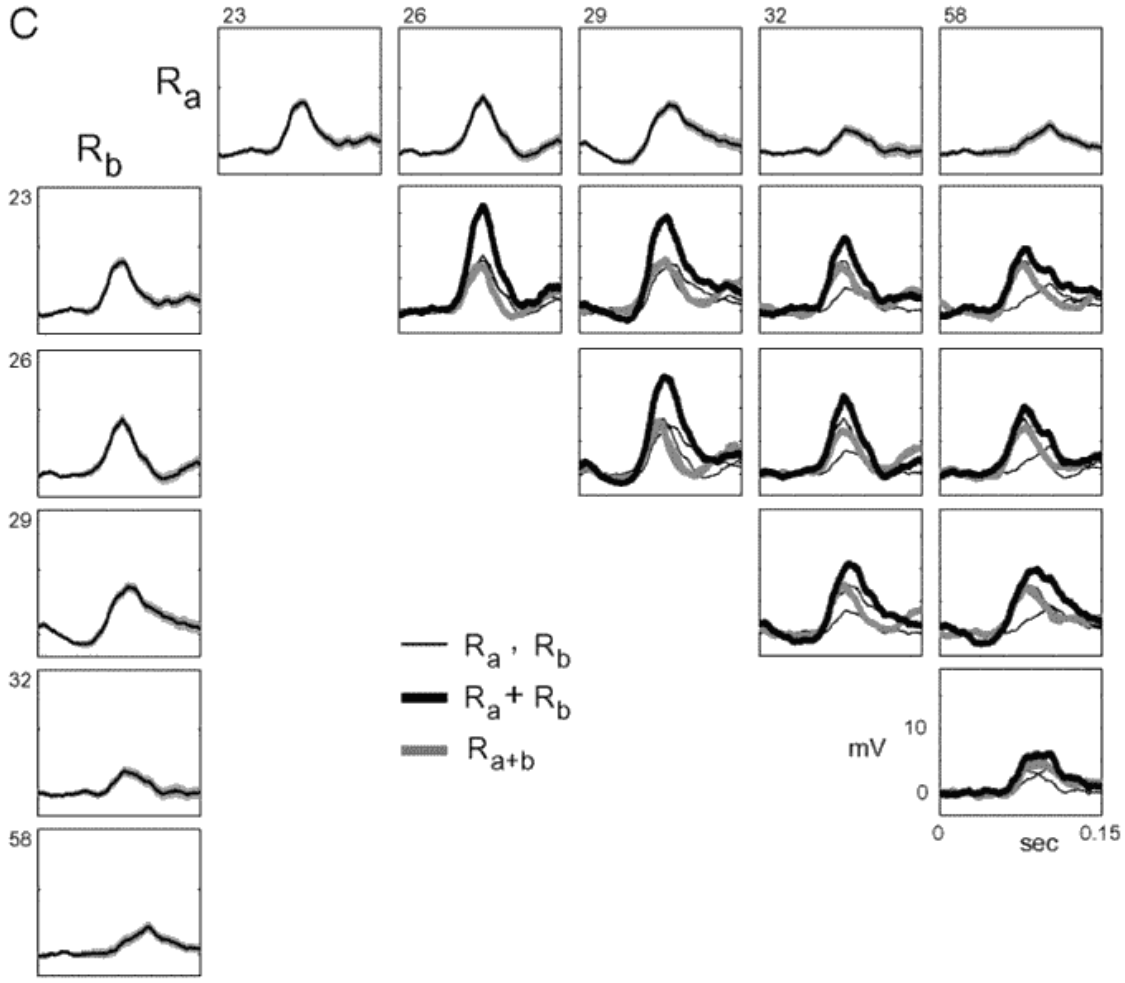


Figure 5.17: Membrane potential response of a cat V1 cell to simultaneous presentation of two bars. [Lampl et al., 2004]

6 Discussion

6.1 A theory of visual cortex

In recent years, we have developed a quantitative and increasingly specific model of the feedforward pathway of the ventral stream in visual cortex – from cortical area V1 to V2 to V4 to IT and PFC – that captures its ability to learn visual tasks, such as identification and categorization of objects from images. The model has been extended during the last years in collaboration with experimental Neuroscience labs. Several of its aspects have been, or are being, tested with extracellular and intracellular recordings in cortex as well as with fMRI and psychophysical techniques.

The quantitative nature of the model has allowed us to directly compare its performance against experimental observations at different scales and also against current computer vision algorithms. At the two extremes of scales that we explored, the model can capture biophysical data and human behavioral data. Specifically, in Section 3 we showed that the model closely matches human performance in a rapid categorization task and in Section 5 we presented biologically plausible circuits that can implement the two main operations of the model and match intracellular voltage recordings in V1 neurons, suggesting a possible canonical microcircuit in cortex [Douglas and Martin, 1991; Poggio and Bizzi, 2004]. In between the biophysics and the psychophysics, we compared the model against electrophysiological recordings at different levels along the ventral visual stream in the macaque visual system. We have shown that (i) the model is consistent with the current knowledge about primary visual cortex (Section 4.1), (ii) the model could well fit and also make predictions about the firing properties in area V4 (Section 4.2), (iii) the model could account for the experimental recordings in IT during presentation of multiple objects (Section 4.3, (iv) the model could explain the observations from the responses in IT upon decoding the population activity to read-out information about object identity and category (see section 4.3).

The model certainly does not account for *all* possible visual phenomena and illusions (see also extensions, predictions and future directions below). However, the success of the model in explaining information across multiple scales and making quantitative predictions strongly suggests that the theory provides an important framework for the investigation of visual cortex. Furthermore, as illustrated throughout this paper, the quantitative implementation of the model allows researchers to be able to perform "experiments" on the model. This is done, for example, by "recording" from units in the model, or making "lesions" in the model.

6.2 No-go results from models

The original motivation for the work on the model was to obtain a computational sanity check about the feasibility of feedforward hierarchical models to account for the main psychophysics and the physiology of the initial phase of visual recognition – before eye movements and attentional shifts.

A strong dissociation between experimental observations and model predictions would suggest that revisions need to be made to the model. This could take the form of psychophysical observations that cannot be explained by the model. Alternatively, this could take the form of two stimuli that yield very different responses in electrophysiological recordings while being indistinguishable for the model and vice versa. This type of contradictory evidence would help refine, upgrade or perhaps even provide no-go results for the theory.

A feedforward architecture from V1 to PFC, very much in the spirit of the Hubel and Wiesel simple-complex hierarchy, seems to account for the main properties of visual cells and in particular for the selectivity and invariance properties of some IT neurons. This was a surprise. In other words it seems that backprojections do not need to play a critical role (apart from possibly priming or "setting up" circuits for a specific recognition task) during the first 150 milliseconds of visual recognition, during which primates can already complete difficult recognition tasks. Notice that a recent experiment measuring information that could be "read-out" from IT cortex (see section 4.3) suggests that after 150 milliseconds from onset of the stimulus performance of the classifier was essentially at its asymptotic performance during passive viewing. Though it is unclear whether this estimate can be generalized to other recognition tasks and different objects and distractors, the data suggest that recognition performance for longer times than 150 milliseconds may be due to activation of the backprojections. In any case, any evidence that feedback plays a key

role in early stages of recognition should be considered as hard evidence that important revisions would need to be made in the architecture of the model.

Of course, the present theory allows for local feedback. In particular, *local* feedbacks circuits within an area may well underlie the two basic operations in the model: tuning via normalization and a max-like operation.

6.3 Extending the theory and open questions

At best, the theory is just a skeleton still missing many important aspects. Here is an incomplete list of the most obvious open questions and predictions:

6.3.1 Open questions

1. How strictly does the hierarchy of Fig. 2.1 map into cells of different visual areas? For instance, are cells corresponding to S2 units in V2 and C2 units in V4 or are some cells corresponding to S2 units already in V1? The theory is rather open about these possibilities: the mapping of Fig. 2.1 is just an educated guess. However, the number of subunits in the cells should increase from V1 to AIT and thus their potential size and complexity. In addition, C units should show more invariance from the bottom to the top of the hierarchy.

2. Do we have sigmoid of dot product or tuning – e.g. sigmoid of normalized dot product – in IT? What before IT? Simulations suggest that they both may work – at least at the level of IT cells.

3. A variety of visual illusions show striking effects that are often counterintuitive and require an explanation in terms of the neuronal circuits. While in some cases specific models have been proposed to explain one or another phenomena, it would be interesting to explore how well the model (and thus feed-forward vision) can account for those observations. A few simple examples include illusory contours (such as the Kanizsa triangle), long-range integration effects (such as the Cornsweet illusion), etc. More generally, it is likely that early Gestalt-like mechanisms – for detecting collinearity, symmetry, parallelism etc. – exist in V1 or V2 or V4. They are not present in this version of the model. It is an open and interesting question how they could be added to it in a plausible way.

4. As stated above, one of the main assumptions of the current model is the feed-forward architecture. This suggests that the model may not perform well in situations that require multiple fixations, eye movements and feedback mechanisms. Recent psychophysical work suggests that performance on dual tasks can provide a diagnostic tool for characterizing tasks that do or do not involve attention [Li et al., 2002]. Can the model perform these dual tasks when psychophysics suggests that attention is or is not required? Are backprojections and feedback required?

5. Are the errors of the model vs. human error in the rapid categorization task of section 3 compatible? Do they suggest deficiencies in the dictionary of features of the model or in its general architecture?

6. The original model, formerly known as HMAX, was extended to deal with recognition of motion and actions [Giese and Poggio, 2003]. Initial work has been done to extend the present theory in the same way [Sigala et al., 2005]. Much more needs to be done. This extension is important because the same IT cells that we discuss here as supporting recognition of static images (the S4, view-tuned units in the model) are likely to be part of networks of reciprocal, lateral, local excitatory connections (learned from passive visual experience) and more global inhibition that endows them with sequence selectivity (see [Sakai and Miyashita, 2004]) and predictivity (Perrett, pers. comm.).

7. Color mechanisms from V1 to IT should be included. The present implementation deals with gray level images. This probably includes a simpler part which involves accounting for the fact that we can recognize the same object under different colors. More complex phenomena involving color such as color constancy and the influence of the background and integration in color perception should ultimately also be explained.

8. Stereo mechanisms from V1 to IT should also be included. Stereo and especially motion also play a important role in the learning of invariances such as position and size invariance via a correlation-based rule such as the trace rule.

6.3.2 Predictions

1. Simple and complex cells have been extensively described and characterized in primary visual cortex. The generic architecture of the model postulates a recurrent connectivity pattern composed of simple cells feeding onto complex cells which show stronger invariance properties. No such clear separation has been clearly established beyond V1 so far.

2. Possible synaptic mechanisms for learning should be described in biophysical details. The theory suggests at least three different synaptic rules for learning correlations at the same time (S layers), for learning correlations across time (C layers) and for task-dependent supervised learning (probably from IT to PFC).

3. The observations in Section 3 and those in 4.3 suggest that accurate object information can be read out from a population of IT neurons in a manner that is robust to the presence of other objects or even in natural complex scenes. The *clear prediction* of the theory is that read-out from "IT" for objects in clutter is possible: the simulations on the animal-no-animal task are with complex natural images with significant clutter. Performance on other databases involving clutter is also very good (see section 3). In particular, we find that the presence of one object can be detected even in the presence of other objects.

4. If blocking GABA A disrupts shunting inhibition, it should disrupts both max and tuning operations (if the biophysical circuits based on shunting inhibition are the correct ones).

6.3.3 Extending the theory to include backprojections

The most critical extension of the theory has to do with the extensive backprojections in visual cortex which need to be taken into account in any complete theory of visual cortex. We sketch the issue in the next section.

In the future, we will have to extend the architecture of our existing model by including backprojections and assigning meaningful functions to them. Our working hypothesis is that a) difficult recognition tasks, as object categorization in complex natural images, can be done within single "snapshots" (*e.g.*, short visual exposures only require the feedforward architecture of the ventral stream), but b) there are recognition tasks (or levels of performance) that need time: such tasks probably require recursions of predictions and verifications (possibly involving eye or attentional "fixations") and the associated engagement of the backprojection pathways.

- There are a number of ideas about the role of backprojections. Backprojections may underlie attentional fixations and zooms-in that may be important in improving performance by focusing on specific spots of the image at the relevant scale and position. In this view, one may try to extend the model to perform visual searches in natural images (i.e. account for eye movements and shifts of attention beyond the first 150 milliseconds). A related proposal to account for task-specific, top-down attention involves the use of a cascade of feedback connections (from PFC to IT and from IT to V4) onto S2 units to bias attention to locations with higher probability of containing a target object (a face, a pedestrian, a car, etc), see [Walther et al., 2005]. A closely related proposal accounts for receptive field dynamics, such as shrinkage and extension. In this possible extension, the C2 pooling range (i.e., the number of S2 units over which the max is taken to compute a C2 response) is a dynamic variable controlled by feedback connections from IT neurons. This could provide a mechanism for computing the approximate object location from the shape pathway.

- A conceptual framework that tries to make sense of the above set of ideas is the following. A program running in PFC decides, depending on the initial feedforward categorization, the next question to ask in order to resolve ambiguity or improve accuracy. Typically, answering this question involves

"zooming in" on a particular subregion of the image at the appropriate level and using appropriate unites (for instance at the S1 level) and calling a specific classifier – out of a repertoire – to provide the answer. This framework involves a flavor of the "20 questions" game and the use of "reverse hierarchy routines" which control access to lower level units.

- Several years ago, a model for translation (and scale) invariant object recognition was put forward in the "shifter" circuit by Anderson and van Essen [Anderson and van Essen, 1987] and was later studied by Olshausen *et al.* [Olshausen et al., 1993] in a system for attention-based object recognition. A routing circuit, putatively controlled by the pulvinar nucleus in the thalamus, was supposed to re-normalize retinal images to fit into a standard frame of reference which was then used for pattern matching to a store of normalized pictures of objects. It is just possible that this model is correct if limited to the description of the attentional mechanisms becoming active *after* the key, initial feedforward categorization step.

- A more sophisticated version of this idea could take the form of a constraint satisfaction network finding an optimal solution satisfying bottom-up constraints (provided by sensory inputs) and top-down constraints (carried by backprojections) after bottom-up and top-down recursions (Geman, pers. comm.). Such a network would have the flavour of currently fashionable graphical models. A similar idea based on factorization of conditional probabilities (or suitable proxies for probabilities) and requiring only a bottom-up sweep (for classification this would be sufficient) and a top-down one (for interpretation) has been suggested by S. Ullman (pers. comm.).

- Could backprojections play a role in top-down segmentation, using the categorization by neurons in PFC to find the S1 cells in the model (simple cells in V1) that were most active?

- Are the backprojections only involved in setting "parameters" in the circuit or in controlling learning?

6.4 A challenge for cortical physiology and cognitive science

- Find another model of recognition that performs as well.

- Alternatively, disprove key aspects of the model.

Acknowledgements We thank Max Riesenhuber, Hiroshi Tsujino, David Ferster, Christof Koch for very useful comments and encouragement. The work was supported by DARPA, ONR, NIH, NSF. We also aknowledge support by Honda, Kodak, Sony, Daimler.

A Appendices

Description of the Appendices

- Appendix A.1 details the model implementation and parameters and shows how the theory evolved from earlier ideas and models.
- Appendix A.2 describes how the tuning properties of the S1 and C1 units were assessed using standard stimuli such as gratings, bars and edges.
- Appendix A.3 provides details about the trace rule learning describe in Section 2.
- Appendix A.4 extends the framework of the theory (2) by comparing different potential fundamental operations that can give rise to tuned neurons, one of the hallmarks of electrophysiological responses in cortical neurons. In particular, the Appendix compares Gaussian tuning, dot products and normalized dot products, and different settings for the p, q and r parameters of Eqn. 1 and Eqn. 2.
- Appendix A.5 argues that the model is robust to changes in many of its parameters. The model is shown to perform well for binary (instead of analog) features.
- Appendix A.6 argues that a linear classifier between IT and PFC is similar to normalized RBF network architecture.
- Appendix A.7 shows that the theory can reproduce recent experiments mapping the activity of V1 neurons in two-spot reverse correlation experiments [Livingstone and Conway, 2003].
- Appendix A.8 extends the observations in 4.2 by elaborating on the fitting technique for the model and on the possible underlying mechanisms for the nonlinear interaction within the receptive fields of V4 and for the relative position tuning.
- Appendix A.9 extends the discussion in 4.3 showing how the theory well accounts for the observations about the fast readout of object information from IT neurons.
- Appendix A.10 shows how PFC can extract information from IT in a task-dependent fashion to perform a task such as categorization Freedman et al. [2001].
- Appendix A.11 presents detailed descriptions of the models and many figures discussed in the text in connection with the biologically plausible models of how the basic operations in the theory can be implemented at the biophysical level (Section 5).
- Appendix A.12 briefly discusses some general questions that are sometimes raised in connection to hierarchical feed-forward models and our model in particular.

A.1 Detailed model implementation and parameters

In this appendix, we provide a detailed model implementation and main parameters. Matlab code for a model subcomponent, $S1 \rightarrow C1 \rightarrow S2b \rightarrow C2b \rightarrow$ *PFC classifier*, corresponding to the bypass routes (yellow lines) in Fig. 2.1 can be found at `http://cbcl.mit.edu/software-datasets/standardmodel/index.html`.

S1 units: The S1 responses are first obtained by applying to the input image I a battery of Gabor filters, which can be described by the following equation:

$$F(u_1, u_2) = \exp\left(-\frac{(\hat{u_1}^2 + \gamma^2 \hat{u_2}^2)}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda}\hat{u_1}\right), \quad \text{s.t.} \tag{10}$$

$$\hat{u_1} = u_1 \cos\theta + u_2 \sin\theta \quad \text{and} \tag{11}$$

$$\hat{u_2} = -u_1 \sin\theta + u_2 \cos\theta, \tag{12}$$

where (u_1, u_2) refers to the S1 receptive field in a 2D coordinate system. Note that we here slightly altered the notation from Section 2 (to emphasize the 2D nature of the S1 receptive fields), *i.e.,* $F(u_1, u_2)$ corresponds to the weight vector $\mathbf{w}$ in the tuning operation (Eq. 1 in Section 2):

$$y = g\left(\frac{\sum_{j=1}^{n} w_j \, x_j^p}{k + \left(\sum_{j=1}^{n} x_j^q\right)^r}\right),$$

In this case the input vector $\mathbf{x}$ corresponds to a small patch of the input image I that falls within the unit's receptive field.

The five parameters (orientation θ, aspect ratio γ, effective width σ, phase ϕ and wavelength λ) determine the properties of the spatial receptive fields. The tuning of simple cells in cortex along these dimensions varies substantially. We consider four orientations ($\theta = 0°, 45°, 90°$ and $135°$). This is an oversimplification but this was previously shown to be sufficient to provide rotation and size invariance at the level of view-tuned units (now S4 units) in good agreement with recordings in IT [Logothetis et al., 1995; Riesenhuber and Poggio, 1999b]. To keep the total number of units tractable, ϕ was set to $0°$ while different phases are crudely approximated by centering receptive fields at all locations.

In order to obtain receptive field sizes consistent with values reported for parafoveal simple cells [Schiller et al., 1976a], we considered 17 filters sizes from 7×7 ($0.19°$ of visual angle) to 39×39 ($1.07°$ of visual angle) obtained by steps of two pixels. When fixing the values of the remaining 3 parameters (γ, λ and σ), we tried to account for general cortical cell properties, that is:

1. Cortical cells' peak frequency selectivities are negatively correlated with their receptive field sizes [Schiller et al., 1976c],
2. Cortical cells' spatial frequency selectivity bandwidths are positively correlated with their receptive field sizes [Schiller et al., 1976c],
3. Cortical cells orientation bandwidths are positively correlated with their receptive field sizes [Schiller et al., 1976b].

We empirically found that one way to account for all three properties was to include fewer cycles with increasing receptive field sizes. We found that the following (*ad hoc*) formulas gave good agreement with the tuning properties of cortical cells:

$$\sigma = 0.0036 \times \textit{RF size}^2 + 0.35 \times \textit{RF size} + 0.18 \tag{13}$$

$$\lambda = \frac{\sigma}{0.8} \tag{14}$$

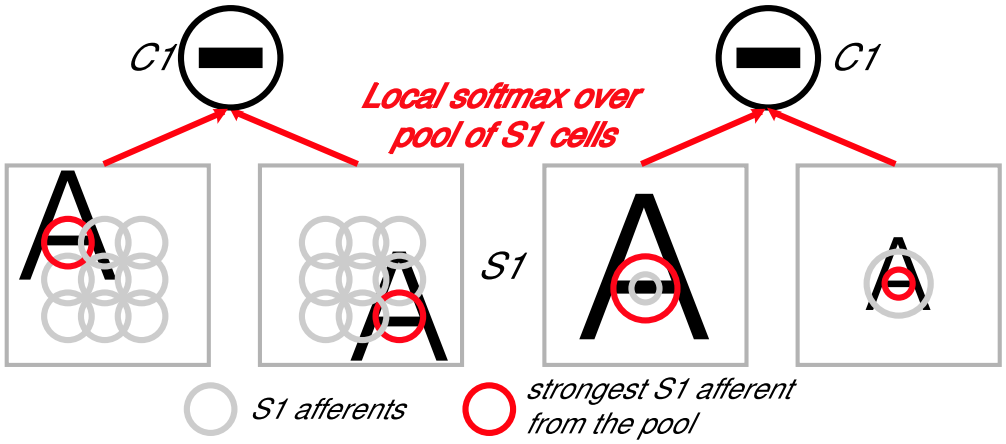


Figure A.1: An example showing how scale and position tolerances are obtained at the C1 level: Each C1 unit receives inputs from S1 units at the same orientation (*e.g.*, 0 degree) arranged in *bands*. For each orientation, a band S contains S1 units in two different sizes and various positions (grid cell of size $N_{C1}^S \times N_{C1}^S$). From each grid cell (see left side) we obtain one measurement by taking the maximum or softmax over all positions: this allows the C1 unit to respond to a horizontal bar anywhere within the grid, providing a translation-tolerant representation. Similarly, taking a max over the two sizes (see right side) enables the C1 unit to be more tolerant to changes in scale.

For all units with a given set of parameters (λ_0, σ_0) to share similar tuning properties at all orientations θ, we applied a circular mask to the Gabor filters. Cropping Gabor filters to a smaller size than their effective length and width, we found that the aspect ratio γ had only a limited effect on the units tuning properties and was fixed to 0.3 for all filters. One can refer to Table 5 for all parameter values. In [Serre and Riesenhuber, 2004] we measured the corresponding S1 units' tuning properties with common stimuli from primary visual cortex (bars, edges and gratings) and showed that they fall within the bulk of parafoveal simple cells. Fig. 4.1 shows the 17×4 S1 units receptive field types.

To summarize, the S1 stage can be obtained by applying a battery of S1 units (Gabor filters) to the input image. The filters come in 4 orientations θ and 17 scales s (see Table 5). This step leads to $17 \times 4 = 68$ maps $(S1)_\theta^s$ that are arranged in 8 bands (e.g., *band 1 contains filters of size 7 and 9, in all 4 orientations).*

C1 units: The next stage – C1 – corresponds to complex cells which show some tolerance to shift and size changes: complex cells tend to have larger receptive fields (twice as large as simple cells on average), respond to oriented bars or edges anywhere within their receptive field [Hubel and Wiesel, 1965b] (shift invariance) and are in general more broadly tuned to spatial frequency than simple cells [Hubel and Wiesel, 1965b] (scale invariance). Position- and scale-tolerances at the C1 level are obtained by pooling over S1 units, as schematically shown in Fig. A.1. Pooling with a softmax-like operation over simple S1 units with the same preferred orientation – but slightly different positions – provides position-tolerance at the C1 level (see Fig. A.1 left). Similarly pooling with a softmax-like operation over simple S1 units at the same preferred orientation – but slightly different scales – provides scale-tolerance at the C1 level (see Fig. A.1 right).

The mathematical form of the operation performed by a C unit is (same as Eq. 3 in Section 2):

$$y = g\left(\frac{\sum_{j=1}^{n} x_j^{q+1}}{k + \left(\sum_{j=1}^{n} x_j^q\right)}\right). \tag{15}$$

In the C layer the input vector $\mathbf{x}$ corresponds to afferent S1 units at the same preferred orientations.

There are three parameters governing this softmax operation. The grid size N_{C1}^S and the scale bands S determine the spatial and size ranges over which inputs are pooled. Note that N_{C1}^S and S determine the C1 unit's receptive field size and the spatial frequency tuning respectively. The parameters in Table 5 were introduced in [Serre and Riesenhuber, 2004] to produce C1 units in good agreements with V1 parafoveal complex cells. Finally to allow for faster processing times, C1 units are down-sampled, *i.e.,* C1 units are computed every ϵ_{C1} S1 units.

To summarize, for each band ($1 \rightarrow 8$), we take a softmax (or max) over scales and positions, i.e., *each band member is sub-sampled by taking the softmax over a grid of size N_{C1}^S* ***and*** *between the two members. For instance for the two S1 maps in band 1, a spatial max is taken over an 8×8 grid* ***and*** *across the two scales (size 7 and 9). Note that we do not take a softmax over different orientations, hence, each band $(C1)^S$ contains 4 orientation maps.*

S2 and C2 units: Each S2 unit receives L ($L = 10$ in our simulations) inputs $(w_{i_1} \dots w_{i_L})$ selected at random from a $M = N_{S2} \times N_{S2} \times 4$ grid of possible afferents $(w_1 \dots w_M)$, where $N_{S2} = 3$ is the spatial range of afferents (thus, the receptive field size of the unit) and 4 is the number of orientations in the previous C1 layer, corresponding to complex cells at different orientations.

Learning occurred in an independent initial training phase in which the model was passively exposed to K_{S2} patches of natural images (see Section 2). The tuning of the S2 units was obtained by setting the weights of each S2 unit-type $w_{i_1} \dots w_{i_L}$ equal to a specific pattern of activity from its C1 afferents in response to a random patch of natural image $\mathbf{x}$ (from a random image at a random position and scale). Recall that each S unit performs a normalized dot-product operation between its inputs $\mathbf{x}$ and its synaptic weights $\mathbf{w}$ (see Eq. 1). Therefore after setting one unit's input weights to the specific response of its C1 afferents to a particular stimulus, the unit becomes *tuned* to this stimulus, which in turn becomes the preferred stimulus of the unit. That is, the unit response is now maximal when a new input $\mathbf{x}$ matches exactly the learned pattern $\mathbf{w}$ and will decrease (with a bell-shape profile) as the new input becomes more dissimilar.

In short when a new input image is presented to the model, each stored S2 unit-type is convolved with the new $(C1)^S$ input image at all scales (this leads to $K_{S2} \times 8$ $(S2)_i^S$ images, where the K_{S2} factor corresponds to the K_{S2} units learned during training and the 8 factor, to the 8 scale bands). S2 units with the same preferred stimuli but at slightly different positions and scales (see parameters governing the pooling in Table 5) are further pooled through a local softmax to obtain $K_{S2} \times 4$ $C2_i^S$ maps, where the 4 comes from the 4 scale bands at the C2 level.

S3 and C3 units: S3 units are computed like S2 units, with $L = 100$ and $N_{S3} = 3$. After taking a final softmax for each $(S3)_i$ map across all scales and positions, we get the final set of K_{S3} shift- and scale-invariant C3 units. The size of our final C3 feature vector thus depends only on the number of S3 unit-types learned during the training stage and not on the input image size.

S2b and C2b units: S2b and C2b units correspond to the bypass routes described in Section 2 (see yellow lines in Fig. 2.1). They are computed like the corresponding S2 and C2 units. Because the receptive fields of the S2b units are larger (see Table 5) and more complex (contain more subunits), they could mimic the kind of tuning that could originate from the direct projections between V2 and PIT.

	S1 parameters							
RF size	7 & 9	11 & 13	15 & 17	19 & 21	23 & 25	27 & 29	31 & 33	35 & 37 & 39
σ	2.8 & 3.6	4.5 & 5.4	6.3 & 7.3	8.2 & 9.2	10.2 & 11.3	12.3 & 13.4	14.6 & 15.8	17.0 & 18.2 & 19.5
λ	3.5 & 4.6	5.6 & 6.8	7.9 & 9.1	10.3 & 11.5	12.7 & 14.1	15.4 & 16.8	18.2 & 19.7	21.2 & 22.8 & 24.4
θ	$0^0; 45^0; 90^0; 180^0$							
	C1 parameters							
Bands S	1	2	3	4	5	6	7	8
grid size N_{C1}^S	8	10	12	14	16	18	20	22
sampling ϵ_{C1}	3	5	7	8	10	12	13	15
	S2 parameters							
num. S2-types K_{S2}	≈ 200							
grid size N_{S2}	3×3 ($\times 4$ orientations)							
num. afferents n_{S2}	10							
	C2 parameters							
Bands S	1 & 2		3 & 4		5 & 6		7 & 8	
grid size N_{C2}^S	8		12		16		20	
sampling ϵ_{C2}	3		7		10		13	
	S3 parameters							
num. S3-types K_{S3}	≈ 500							
grid size N_{S3}	3×3 ($\times K_{S2}$)							
num. afferents n_{S3}	100							
	C3 parameters							
Bands S	1 & 2 & 3 & 4 & 5 & 6 & 7 & 8							
grid size N_{C3}^S	40							
	S2b parameters							
num. S2b-types K_{S2b}	$\approx 500 \times 4$ (for each patch size)							
patch size P_{S2b}	$6 \times 6; 9 \times 9; 12 \times 12; 15 \times 15$ ($\times 4$ orientations)							
num. afferents n_{S2b}	100							
	C2b parameters							
Bands S	1 & 2 & 3 & 4 & 5 & 6 & 7 & 8							
grid size N_{C2b}^S	40							

Table 5: Summary of all model parameters (see accompanying text).

How the new version of the model evolved from the original one The architecture sketched in Fig. 2.1 has evolved – as originally planned and from the interaction with experimental labs – from the original one described in [Riesenhuber and Poggio, 1999b]. In particular new layers have been added to improve the mapping between the functional primitives of the theory and the structural primitives of the ventral stream in the primate visual system. Below is a list of major changes and differences between this new model implementation and the original one:

1. **The two key operations:** Operations for selectivity and invariance, originally computed in a simplified and idealized form (*i.e.*, a multivariate Gaussian and an exact max, see Section 2) have been replaced by more plausible operations, normalized dot-product and softmax (see Section 2 and Section 5 for biophysically-plausible circuits).

2. **S1 and C1 layers:** In [Serre and Riesenhuber, 2004] we found that the S1 and C1 units in the original model were too broadly tuned to orientation and spatial frequency and revised these units accordingly. . In particular at the S1 level, we replaced Gaussian derivatives with Gabor filters to better fit parafoveal simple cells' tuning properties. We also modified both S1 and C1 receptive field sizes.

3. **S2 layers:** They are now learned from natural images (see Section 2 and Appendix A.3). S2 units are more complex than the old ones (simple 2×2 combinations of orientations). *The introduction of learning, we believe, has been the key factor for the model to achieve a high-level of performance on natural images, see [Serre et al., 2002].*

4. **C2 layers:** Their receptive field sizes, as well as range of invariances to scale and position have been decreased so that C2 units now better fit V4 data. See Section 4.2.

5. **S3 and C3 layers:** They were recently added and constitute the top-most layers of the model along with the S2b and C2b units (see Section 2 and above). The tuning of the S3 units is also learned from natural images.

6. **S2b and C2b layers:** We added those two layers to account for the bypass route (that projects directly from V1/V2 to PIT, thus bypassing V4 [see Nakamura et al., 1993]). Interestingly these bypass routes have been shown to provide an excellent compromise (when used alone) between speed and accuracy in computer vision applications (see [Serre et al., 2005c,b]). Of course, units from the bypass route in the model could also be provided less directly through a relay stage in V4.

A.2 Comparing S1 and C1 units with V1 parafoveal cells

This subsection describes how the tuning properties of the S1 and C1 units were assessed using standard stimuli such as gratings, bars and edges. The reader may refer to [Serre et al., 2004] for supplementary information and how the units in the new implementation differ from the original one [Riesenhuber and Poggio, 1999b].

A.2.1 Methods

Orientation tuning The orientation tuning of a model unit was assessed in two ways: First, following [Valois et al., 1982b], we swept a sine wave grating of optimal frequency over the receptive field of the unit at thirty-six different orientations (spanning 180^o of the visual field in steps of 5^o). For each cell and orientation tested, we recorded the maximum response (across positions) to further fit a tuning curve and compute the orientation bandwidth at half-amplitude. For comparison with [Schiller et al., 1976b], we also swept edges and bars of optimal dimensions: For each cell the orientation bandwidth at 71% of the maximal response was calculated as in [Schiller et al., 1976b].[1]

Spatial frequency tuning The spatial frequency selectivity of a model unit was assessed by sweeping sine wave gratings of various spatial frequencies over the receptive field of the unit. For each grating frequency, the maximal cell response was recorded to fit a tuning curve and the spatial frequency selectivity bandwidth was calculated as in [Valois et al., 1982a] by dividing the frequency score at the high crossover of the curve at half-amplitude by the low crossover at the same level.

Taking the $\log_2$ of this ratio gives the bandwidth value (in octaves):

$$\text{bandwidth } = \log_2 \frac{\text{high cut}}{\text{low cut}} \tag{16}$$

For comparison with [Schiller et al., 1976c], we also calculated the *selectivity index* as defined in [Schiller et al., 1976c], by dividing the frequency score at the high crossover of the curve at 71% of the maximal amplitude by the low crossover at the same level and multiplying this value by 100 (a value of 50 representing a specificity of 1 octave):

$$\text{selectivity index} = \frac{\text{high cut}}{\text{low cut}} \times 100 \tag{17}$$

A.2.2 Spatial frequency tuning

S1 units. Gabor filter peak frequencies are parameterized by the inverse of their wavelength $\nu = \frac{1}{\lambda}$ (*i.e.*, the wavelength of the modulating sinusoid, see Appendix A.1). We found that the values measured experimentally by sweeping optimally oriented gratings were indeed close to ν. As expected (see Appendix A.1), we also found a positive correlation between receptive field size and frequency bandwidth, as well as a negative correlation with peak frequency selectivities, which is consistent with recordings made in primate striate cortex [Valois et al., 1982a; Schiller et al., 1976c].

The peak frequency of model units was in the range 1.6-9.8 cycles/degree (mean and median value of 3.7 and 2.8 cycles/degree respectively). This provides a reasonable fit with cortical simple cells peak frequencies lying between values as extreme as 0.5 and 8.0 degree/cycles but a bulk around 1.0-4.0 cycles/degree (mean value of 2.2 cycles/degree) [Valois et al., 1982a]. Indeed, using our formula to parameterize Gabor filters (see Appendix A.1), a cell with a peak frequency around 0.5 cycles/degree would have a receptive field size of about 2^o which is very large compared to values reported in [Hubel and Wiesel, 1965a; Schiller et al., 1976a] for simple cells.

Spatial frequency bandwidths measured at half-amplitude were all in the range 1.1-1.8 octaves, which corresponds to a subset of the range exhibited by cortical simple cells (values reported as extreme as 0.4 to values above 2.6 octaves). For the sake of simplicity, we tried to capture the bulk of "frequency bandwidths" (1-1.5 octaves for parafoveal cells) and focused on population median values (1.45 for both cortical [Valois et al., 1982a]) and model cells). For comparison with Schiller *et al.* , we measured the spatial frequency index and found values in the range 44-58 (median 55) which lies right in the bulk (40-70) reported in [Schiller et al., 1976c].

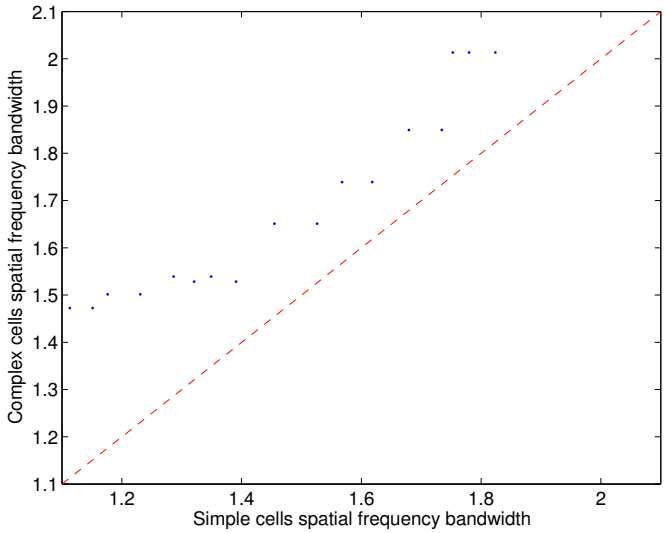


Figure A.2: Complex cells *vs.* simple cells spatial frequency bandwidth. There is an increase of about 20% from simple to complex cells spatial frequency bandwidth, consistent with parafoveal cortical cells [Schiller et al., 1976c; Valois et al., 1982a].

C1 units. The peak frequencies ranged from 1.8 to 7.8 cycles/degree (mean value and median values of 3.9 and 3.2 respectively) for our model complex cells. In [Valois et al., 1982a], peak frequencies range between values as extreme as 0.5 and 8 cycles/degree with a bulk of cells lying between 2-5.6 cycles/degree (mean around 3.2).

We found the spatial frequency bandwidths at half-amplitude to be in the range 1.5-2.0 octaves. Parafoveal complex cells lie between values as extreme as 0.4 to values above 2.6 octaves. Again, we tried to capture the bulk frequency bandwidths ranging between 1.0 and 2.0 octaves and matched the median values for the populations of model and cortical cells [Valois et al., 1982a] (1.6 octaves for both). The spatial frequency bandwidth at 71% maximal response were in the range 40-50 (median 48) which lies within the bulk (40-60) reported in [Schiller et al., 1976c]. Fig A.2 shows the complex *vs.* simple cells spatial frequency bandwidths.

A.2.3 Orientation tuning

S1 units. We found a median orientation bandwidth at half amplitude of 44^o (range 38^o-49^o). In [Valois et al., 1982b], a median value of 34^o is reported. Again, as already mentioned earlier, this value seems surprising (it would imply that parafoveal cells are more tightly tuned than their foveal homologue, both simple (median value 42^o) and complex (45^o). When we used instead a measure of the bandwidth at 71% of the maximal response for comparison with Schiller *et al.* , the fit was better with a median value of 30^o (range: $27^o - 33^o$) compared with a bulk of cortical simple cells within $20^o - 70^o$[Schiller et al., 1976b].

C1 units. We found a median orientation bandwidth at half amplitude of 43^o which is in excellent agreement with the 44^o reported in [Valois et al., 1982b]. The bulk of cells reported in [Schiller et al., 1976b] is within $20^o - 90^o$ and our values range between $27^o - 33^o$ (median 31^o), therefore placing our model units as part of the most narrowly tuned subpopulation of cortical complex cells. As in both experimental data sets, the orientation tuning bandwidth of the model complex units is very similar to that of simple units.

Notes

[1]Sweeping edges, bars and gratings gave very similar tuning curves.

A.3 Training the model to become an expert

While we found in Section 3 that a universal dictionary of features (learned from random patches of natural images) was sufficient to enable the model to perform many different recognition tasks, we found that the model performance could be improved by over-training the model to become an expert. Interestingly training the model to become an expert requires: 1) to present the model with more training examples and 2) longer training times. This step consists in selecting a subset of the *S* units that are selective for a particular target-object class (*e.g.,* face, car or animal) among the very large set of all *S* units learned from natural images. Again it is important to point out that this selection step is optional and in brains probably only occurs for few object classes for which we are experts (*e.g.,* faces). This step is applied to all S2, S2b and S3 units so that, at the top of the hierarchy, the S4 units corresponding to view-tuned units in IT receive inputs from object-selective afferents.

Learning in visual cortex To-date little is known about learning and how visual experience shapes the tuning properties of neurons in cortex. At the computational level various clues or principles of natural vision have been proposed, which could potentially be exploited by a biological organism to guide learning, *e.g.,* Barlow's suspicious coincidences [Barlow, 1989] or smoothness properties [Sutton and Barto, 1981; Marr, 1982; Becker and Hinton, 1992; Stone and Bray, 1995; de Sa and Ballard, 1998], *etc.*

Related work Recently several algorithms for learning transformation-sequences from temporal association have been proposed. They use a trace learning rule to exploit the temporal smoothness of an image sequence that contains an object under transformations (*e.g.,* a head rotating to learn pose-invariance, an object moving in the 2-D plane to learn shift-invariance or an object looming for scale-invariance). Such an algorithm was originally proposed by Sutton & Barto for classical conditioning [Sutton and Barto, 1981] and later for invariance to pose [Perrett et al., 1984, 1985], translation [Földiák, 1991; Einhäuser et al., 2002; Wiskott and Sejnowski, 2002; Spratling, 2005], light [Hietanen et al., 1992] and occlusion of object parts [Wachsmuth et al., 1994]. As a plausibility proof a trace learning rule was also successfully implemented in VisNet [Wallis et al., 1993; Wallis and Rolls, 1997; Elliffe et al., 2002], a model of invariant object recognition (see [Deco and Rolls, 2004] for a recent review).

The Algorithm The proposed approach seeks *good* units to represent the target-object class, *i.e.,* units that are robust to target-object transformations (*e.g.,* inter-individuals variations, lighting, pose, *etc.*). A detailed presentation of the learning algorithm is provided in Fig. A.3 and Fig. A.4. By presenting the model with sequences of images that contain the target-object embedded in various backgrounds (see Fig. 3.2 for typical stimuli used), the trace rule learning algorithm (*TR*) selects features that are robust against clutter and within-class shape variations. Beyond shape recognition in the ventral pathway, we have successfully extended the *TR* algorithm to the recognition of biological motion in a model of the dorsal pathway [Sigala et al., 2005].

Evidence for the use of temporal association during learning Behavioral evidence for temporal association of view comes from a study by Wallis & Bülthoff. They showed that faces from different people that are presented during training as a continuous image sequence of a rotating face will later be associated as being from the same person [Wallis and Bülthoff, 2001]. Interestingly such memory *trace* could be easily implemented in cortex, for instance in the maintained firing rate of neurons in IT after stimulus offset for several hundreds of milliseconds or indeed any kind of short-term memory [Deco and Rolls, 2004; Rolls and Tovee, 1994]. The notion of a temporal association between a sequence of inputs seems consistent – as pointed out by Stryker [Stryker, 1991; Földiák, 1998; Giese and Poggio, 2003] – with a study by Myashita, who showed, that training a monkey with a fixed sequence of image patterns lead to a correlated activity between those same patterns during the delayed activity [Miyashita, 1988].

Initialization: The algorithm starts with the random selection of an initial pool of K_{S2} S2 units from the very large *universal* dictionary of (S2) units that are tuned to patches of natural images. Each unit i in the pool is denoted $\mathbf{w^i} = (w_1^i \dots w_n^i)$. Its response to a small image patch $\mathbf{x}$ falling inside its receptive field is given by Eq. 1 and is denoted $\mathbf{S_{w^i}(x)}$. Each unit i is associated with a fitness function ϕ_i, called (memory) trace because it characterizes the recent activity of the unit (see below).

Iterations: Units are presented with a sequence of images that contain the target-object (*e.g.,* faces). For image t in the sequence, K random patches $\mathbf{x_k}$ are extracted and presented to all M units in the pool simultaneously. Key in the algorithm is that *units compete for reward*, *i.e.,* the trace of each unit $\phi_i(t)$ follows an exponential decay α and will only be increased – by an amount $\beta_i(t)$ – if the unit has won at least once for the current image presentation t (*i.e.,* has had the max response across the pool for at least one image patch). Because units *die* (and are replaced at random) whenever their traces fall below a fixed threshold θ, the only way for a unit to survive is to win as often as possible. For each image presentation t, this corresponds to the following update:

$$
\begin{aligned}
I(t) &= \{\text{argmax}\left(\mathbf{S_{w^i}(x^k(t))}\right)\}_{k=1\dots K} \\
if \quad i \in I & \qquad \beta_i(t) = 1 \\
else & \qquad \beta_i(t) = 0 \\
\phi_i(t) &= \alpha\beta_i(t-1) + \beta_i(t), \text{ where } \alpha < 1 \\
if \quad \phi_i(t) &< \theta, \text{ replace unit } \mathbf{w^i} \text{ at random.}
\end{aligned}
$$

Figure A.3: The trace rule (*TR*) learning algorithm.

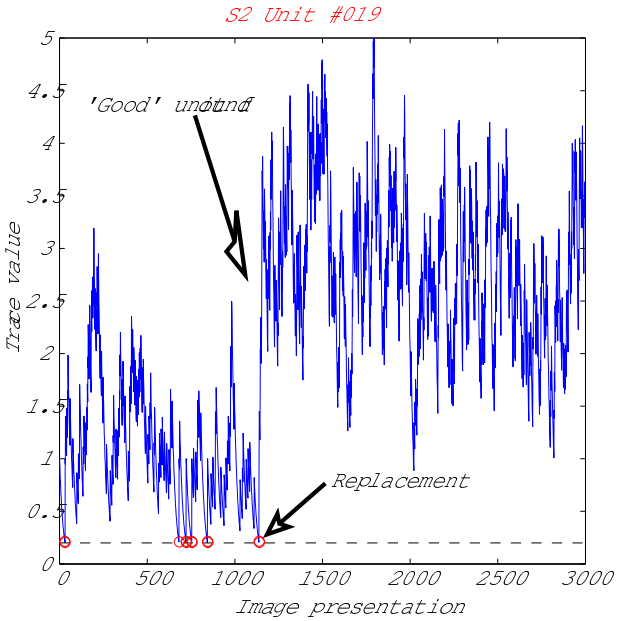


Figure A.4: Tracking a S2 unit (#19) during learning with the trace rule (*TR*) algorithm. Here the model was passively exposed to a sequence of images containing face examples: The trace value of the unit (y axis) varies as a function of time (x axis) as images are presented in sequence. When the unit is "doing well" it is being rewarded (*i.e.,* its trace increases), else it is punished (its trace decreases). Whenever the trace of the unit goes below a fixed threshold (indicated by a dash line) the unit dies and is being replaced by a new one (tuned to a different patch of natural image). As one can see when a good unit is picked (around presentation #1,100 on this example), its trace goes high and remains stable until the algorithm converges.

A.4 Comparison between Gaussian tuning, normalized dot product and dot product

A.4.1 Introduction

Across the cortex, especially in the sensory areas, many neurons respond strongly to some stimuli, but weakly to others, as if they were tuned to some optimal features or to particular input patterns. For example, neurons in primary visual cortex show Gaussian-like tuning in multiple dimensions, such as orientation, spatial frequency, direction, and velocity. Moving further along the ventral pathway of primate cortex, V4 neurons show tuned responses to different types of gratings or contour features [Gallant et al., 1996; Pasupathy and Connor, 2001], and some IT neurons are responsive to a particular view of a face or other objects [Logothetis et al., 1995; Kobatake and Tanaka, 1994]. The tuning of a neuron may be sharp and sparse in some cases, or distributed and general in other cases [Kreiman, 2004], but despite qualitative differences, such tuning behavior seems to be one of the major computational strategies for representing and encoding information in cortex.

Even though Gaussian-like tuning behavior in cortex is widely acknowledged, it remains a major puzzle: how could such multidimensional tuning be implemented by neurons? The underlying biophysical mechanism is not understood. In Hubel and Wiesel's model of V1, the tuning properties of simple and complex cells are explained in terms of the geometry of the afferents: for simple cells, the alignment of several non-oriented LGN afferents would give rise to the orientation selectivity (see [Ferster and Miller, 2000] for a review, and [Ringach, 2004a] for a quantitative model). Although attractively simple and intuitive, this explanation is challenged by a competing theory that maintains orientation selectivity is enforced, if not created, by the recurrent neural circuitry within V1 [Somers et al., 1995; Ben-Yishai et al., 1995]. The tuning along non-spatial dimensions such as velocity or color, however, cannot rely on the geometric arrangements only. Furthermore, tuning in other sensory modalities (*e.g.,* auditory or olfactory neurons) and in higher visual areas where the tuning seems to be of a more abstract nature (*e.g.,* the complex shape tuning in IT) would require a more general mechanism.

Our proposal for tuning operation, as defined by Eq. 1, is based on divisive normalization, weighted sum, and sigmoid-like nonlinearity, all of which are biologically plausible. This operation implies that the tuning emerges from feedforward inputs *and* intra-cortical interactions. For example, in the case of V1, the orientation selectivity is a result of not just geometrical layout of the LGN inputs, but also intracortical interaction that sharpens the tuning. The advantage of considering normalized dot product as a tuning operation is its biophysical plausibility. Unlike the computation of Euclidean distance or a Gaussian function, both normalization and dot product operations can be readily implemented with a network of neurons. The dot product or the weighted sum can be computed by the dendritic inputs to a cell with different synaptic weights. The normalization across the inputs can be achieved by a divisive gain control mechanism involving inhibitory interactions [Reichardt et al., 1983; Carandini and Heeger, 1994; Carandini et al., 1997; Heeger, 1993]. The sigmoid-like scaling of neural responses, possibly nonlinearities in axon or soma, would control the sharpness of the tuning.

A.4.2 Normalized dot product *vs.* Gaussian

Consider the following mathematical identity, which relates the Euclidean distance measure with the normalized dot product:

$$|\vec{x}-\vec{w}|^2 = -2\vec{x}\cdot\vec{w}+1+|\vec{w}|^2, \text{if } |\vec{x}|=1. \tag{18}$$

In other words, the similarity between two normalized vectors, $\vec{x}$ and $\vec{w}$, can be measured with a Euclidean distance as well as a dot product, or the angle between two vectors. Hence, Eq. 18 suggests that Gaussian function $f(\vec{x}) = e^{|\vec{x}-\vec{w}|^2/2\sigma^2}$, which is based on $|\vec{x}-\vec{w}|^2$, is very closely related to a normalized dot product operation and can be approximated by it. Here we quantitatively compare a Gaussian tuning function and a normalized scalar product (NSP) with a sigmoid nonlinearity:

$$R_{Gauss} = e^{|\vec{x}-\vec{w}|^2/2\sigma^2}, \tag{19}$$

$$R_{NSP} = \frac{1}{1+e^{-\alpha(\frac{\vec{x}\cdot\vec{w}}{|\vec{x}|+k}-\beta)}}. \tag{20}$$

The sigmoid is a commonly used transfer function for modeling the relationship between the pre-synaptic and post-synaptic activations or membrane depolarizations in neurons. It sharpens the tuning behavior

created by normalized dot product and allows a better approximation of the Gaussian function, as the parameters α and β are adjusted. For R_{Gauss}, $\vec{w}$ specifies the center of the Gaussian function in a multi-dimensional space. The Gaussian width σ determines the sharpness or sensitivity of tuning (σ need not be the same along different dimensions). For R_{NSP}, $\vec{w}$ specifies the direction of the feature vector along which the response is maximal, and the parameters α and β determine the sharpness of the tuning. In both cases, the response is maximal if the input $\vec{x}$ is matched to the target $\vec{w}$. Fig. A.5 shows a few direct comparisons between R_{Gauss} and R_{NSP}. Although not identical, R_{NSP} and R_{Gauss} exhibit comparable tuning behaviors.

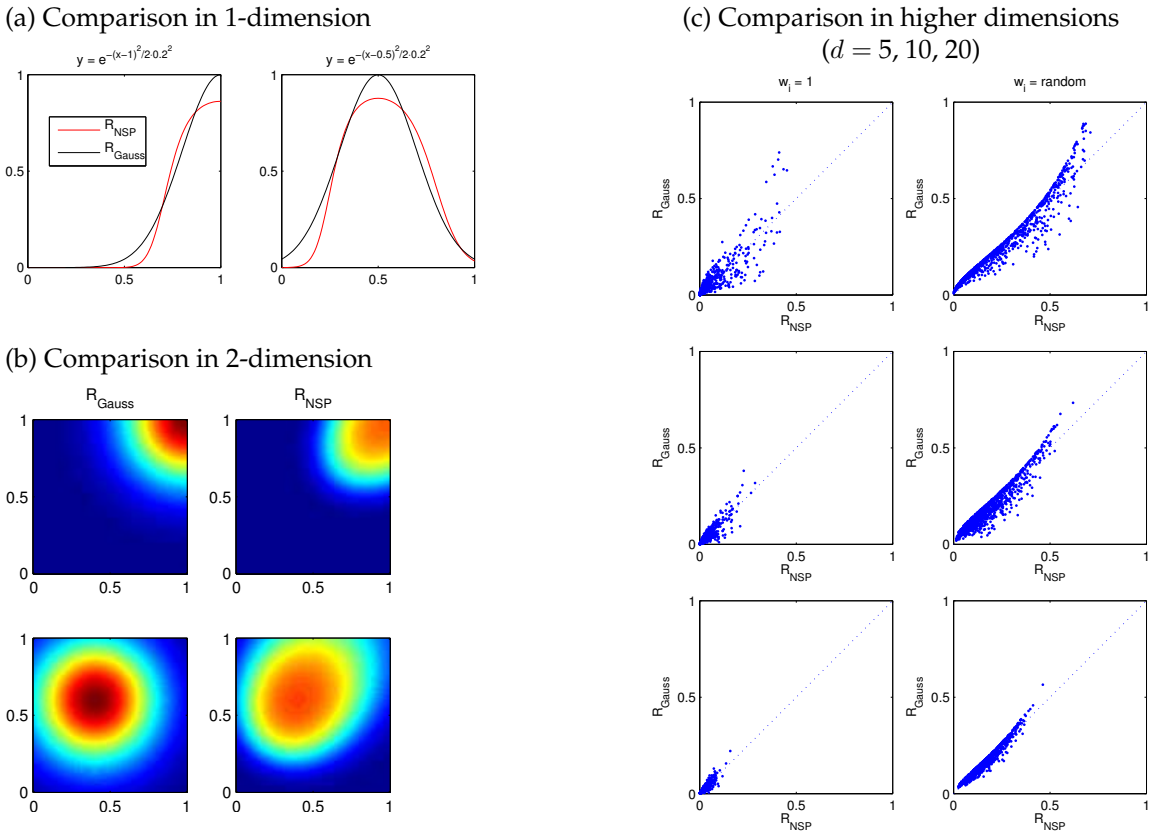


Figure A.5: Comparison of R_{Gauss} and R_{NSP} in several dimensions. Note that in all cases, the Gaussian tuning function can be approximated by the normalized dot product followed by a sigmoid nonlinearity. The parameters α and β in the sigmoid are found with nonlinear fitting (`nlinfit` in `Matlab`), while k was fixed at 0.1. As pointed out in the Remarks section, a dummy input (a bias term) was introduced to obtain tuning to an arbitrary $\vec{w}$ (*i.e.*, R_{NSP} is in $\mathcal{S}^{n+1}$). (a) Comparison in 1-dimension: R_{Gauss} (black) with $\sigma = 0.2$ and R_{NSP} (red) are shown for $w = 1$ (left) and $w = 0.5$ (right). (b) Similar comparisons in 2-dimension: $\vec{w} = (1, 1)$ (top) and $\vec{w} = (0.4, 0.6)$ (bottom). (c) Comparisons in higher dimensions. Since the visualization of the entire function is difficult for high dimensions, 1000 random points are sampled from the space. The same nonlinear fitting routine was used to find the parameters in R_{NSP}. The width of Gaussian is scaled according to $\sigma = 0.2\sqrt{d}$, where d is the dimensionality.

A.4.3 Can a tuning behavior be obtained for $p \simeq q$ and $r = 1$?

Fig. A.6 shows the behavior of $y = \frac{\sum x_i^p}{c + \sum x_i^q}$, when $(p, q) = (1, 1.2)$ or $(0.8, 1)$ in one dimension. As long as $p < q$, at large values of x, y decreases, and by choosing the sigmoid function appropriately, we can make a tuning behavior. The location of the peak can also be controlled with the parameters in the function.

For $p = 1$, $q = 2$ and $r = 1/2$ (or $p = q = r = 1$), the normalized dot product only increases with x. Therefore, a bell-shaped tuning about an arbitrary center is not possible, unless a fixed bias term, such as the constant leak conductance, is assumed (see Remarks).

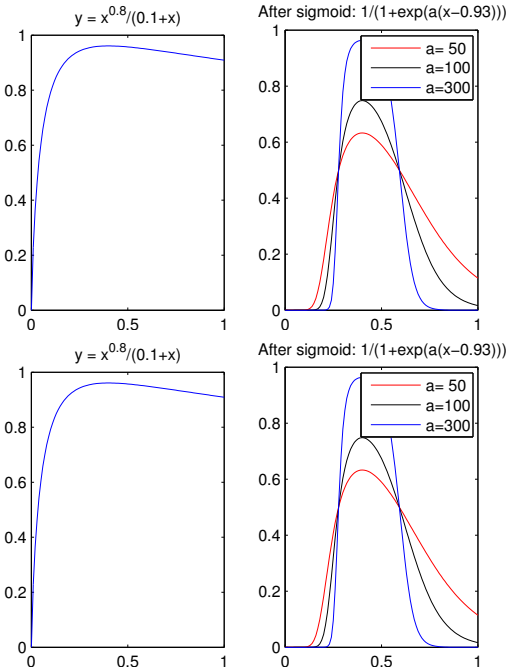


Figure A.6: Two examples of tuning. Left: Normalized dot product only. Right: Normalized scale product followed by sigmoid.

A.4.4 Dot product *vs.* normalized dot product *vs.* Gaussian

How much nonlinearities does the model require to function properly. In case of the tuning or the template matching nonlinearity,

$$\text{Linear: } \vec{x} \cdot \vec{w} \leftrightarrow \text{Nonlinear: } \frac{\vec{x}}{|\vec{x}|} \cdot \vec{w} \leftrightarrow \text{Gaussian: } e^{-(\vec{x}-\vec{w})^2}. \tag{21}$$

As the normalized scalar product with sigmoid can approximate the Gaussian function well, this operation can be used to explore the degree of nonlinearity. For now (without giving much thought on biophysical mechanism yet), consider the following toy example:

$$y = \frac{\vec{x} \cdot \vec{w}}{|\vec{x}|^n}, \tag{22}$$

where $0 \leq n \leq 1$. When $n = 0$ this operation is purely linear, and when $n = 1$ it is the exactly the normalized scalar product.

Now imagine a neuron whose output response to the input pattern is determined by the above equation. $\vec{w}$ is a target vector to which the neuron is tuned. It will be the synaptic weights in terms of the normalized scalar product or the center of the Gaussian function. Consider a benchmark test like the paperclip experiment. Assume that the number of afferent inputs are fixed (*i.e.,* input dimension = 10, 25, and 100), and consider how much invariance and selectivity against distractors can be measured using different degrees of tuning nonlinearity. From the onset, we would expect that simple linear operation would give worse performance than the highly nonlinear operation.

Note that the normalized scalar product was able to produce similar selectivity as the Gaussian function. As the normalization is weakened and the operation gets more linear, the selectivity decreases (*i.e.,* the distractor responses are comparable to or even higher than the tuned responses).

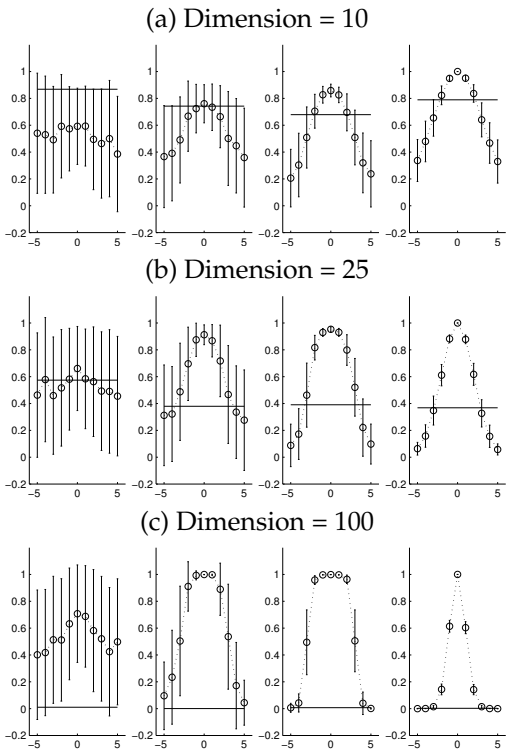


Figure A.7: We looked at the tuned response about a random target vector $\vec{w}$, by considering $\vec{x} = \vec{w} + \alpha$, where α goes from 0 to some finite values. (In terms of the paperclip benchmark, α could represent the viewpoint rotation or scale differences from the tuned paperclip image.) The responses to distractors are created by considering random $\vec{x}$ and computing the normalized scalar product with $\vec{w}$. By comparing with the maximum response to the distractors, we can measure the amount of selectivity and invariance, against distractors. Sigmoid function was applied to scale the responses between 0 and 1. From top to bottom, we used different numbers of afferent inputs. From left to right, $n = 0.1, 0.5, 0.9$ and far right is the Gaussian tuning function for comparison. The horizontal line represents the maximum response to the 100 distractors. The tuning curves are quite variable (especially when the response function is closer to the linear operation. These results are the averages and standard deviations of 100 trials. The normalized scalar product considered in these simulations did not have the dummy input (see the Remarks). Also note that the normalized scalar product shows flatter tuning curve than the Gaussian, as pointed out in [Maruyama et al., 1992].

However, when the number of afferents is large (high input dimension or more synapses), simple scalar product seems to work to a certain degree (*i.e.,* the tuned response was greater than the maximum distractor response in some, not all, cases), although the performance is much poorer. In other words, when the input dimension is quite large, the input and the target vectors are seldom aligned, and therefore, there is little need to normalize the input vector. The implication is that the elaborate normalization scheme may not be even necessary, if the neuron can listen to many afferents.

Most likely, there will be some trade-offs. *The crudely tuned response can be created by the scalar product (*i.e., *the connectivity between the afferents and the output cell), and the tuning is refined by the normalization circuit. Interestingly, this is somewhat analogous to the V1 case, where the orientation selectivity is believed to be initially determined by the connectivity with LGN cells and the tuning gets sharper with recurrent connections in V1.*

This tradeoff may be that of efficiency. It will be easy, but not precise to accomplish tuning with many, many features. On the other hand, the precise and economical (in terms of the number of afferent cells and maintaining the active synapses) to used more nonlinear tuning function, but this nonlinearity will also be subject to approximations and noises. Hence, a larger number of afferents can compensate for the lack of "tuned-ness" or nonlinearities like the divisive normalization.

Notes

[2]This relationship between Gaussian and normalized dot product was pointed out by Maruyama, Girosi and Poggio in [Maruyama et al., 1992], where the connection between the multilayer perceptron and the neural network with radial basis function is explored. Their analysis is based on the exact form of this identity (*i.e.,* the input $\vec{x}$ to the Euclidean distance is normalized as well as the input to the dot product). In this paper, we examine a looser connection between the Euclidean distance and the normalized dot product (*i.e.,* the input to the Euclidean distance is not, but the input to the dot product is normalized):

$$|\vec{x} - \vec{w}|^2 \leftrightarrow \frac{\vec{x} \cdot \vec{w}}{|\vec{x}|}.$$

[3]Since the dot product $\vec{x} \cdot \vec{w}$ measures the cosine of the angle between two vectors, the maximum occurs when those two vectors are parallel. Because it is also proportional to the length of the vector, a simple dot product is not as flexible as Gaussian function which can have an arbitrary center. However, a simple workaround is to assume a constant dummy input, which introduces an extra dimension and allows tuning for any $\vec{w}$. In other words, because of the normalization, the dimensionality of R_{NSP} is one less than that of R_{Gauss}. With the same number of afferents n, the Gaussian tuning function may be centered at any points in $\mathcal{R}^n$, whereas the normalized dot product is tuned to the direction of the vector in $\mathcal{S}^n$ or $\mathcal{R}^{n-1}$. An obvious way of avoiding such limitation is to assume a constant dummy input (a fixed bias term, which may originate from leak conductance or the resting activity of a neuron) and to increase the dimensionality of the input vector, which was the approach taken here. Then, the normalized dot product may be tuned to any arbitrary vector $\vec{w}$, just like the Gaussian function. See [Maruyama et al., 1992; Kouh and Poggio, 2004].

[4]The normalization for tuning provides some new insights and predictions. For example, along the ventral pathway of primate visual cortex, the receptive field size on average increases, and neurons show tuning to increasingly complex features [Kobatake and Tanaka, 1994]. In order to build a larger receptive field and to increase feature complexity, the neurons may be pooling from many afferents covering different parts of receptive fields. The afferent cells within the pool would interact via normalization operation, whose interaction may appear as a center-surround effect as observed in V1 [Cavanaugh et al., 2002]. If indeed a general mechanism for tuning, normalization would be present in other cortical areas, where similar center-surround or interference effects may be observable. The effects of normalization may also appear whenever the response of one afferent in the normalization pool is modulated (for example, an attentional mechanism through a feedback connection). Change in one neuron's response may affect not only the output of the network, but also the response of other afferent neurons in the normalization pool.

[5]Note that this scheme for cortical tuning has implications for learning and memory, which would be accomplished by adjusting the synaptic weights according to the activation patterns of the afferent cells.

[6]In the past, various neural microcircuits have been proposed to implement a normalization operation. The motivation was to account for gain control. We make here the new proposal that another role for normalizing local circuits in brain is to provide the key step for multidimensional, Gaussian-like tuning. In fact this may be the main reason for the widespread presence of gain control circuits in cortex where tuning to optimal stimuli is a common property.

[7]If we further assume that the normalization operation kicks in later than the scalar product (possibly due to the delay of going through the pool cell or feedback inhibition), then the nonlinear, sharply tuned response will appear later than the linear response.

A.5 Robustness of the model

In this appendix we provide a complementary analysis for Section 2 and describe the model dependency on 1) the use of analog *vs.* binary values at the level of the S4 afferents and 2) the tuning operation (*i.e.,* exact Gaussian tuning, normalized dot-product and simple dot-product).

Our analysis relies on two measures of the model performance, *i.e.,* the model categorization performance on real-world natural image datasets and the invariance properties of units at the S4 (VTUs) level against the benchmark paperclip dataset [Logothetis et al., 1995].

As discussed in Section 5, we speculate that *"the cable may become thinner and thinner"* along the hierarchy. Here we take an extreme approach and evaluate the model performance when units response in higher stages (inputs to S4) are binarized, *i.e.,* units become like switches that are either *on* or *off*).

To evaluate the model performance on natural images we considered a subset of the 101-object dataset [Fei-Fei et al., 2004] (see Section 3). We first selected the image categories that contained at least 150 examples so as to generate random splits (100 training, 50 test cross-validation). This included five sets (faces, leopards, motorcycles, airplanes, watches) plus a "background" category (target absent).

We tested the model on two recognition problems: a binary class classification problem (target object present or absent, chance level: 50%) and a multi-class classification problem with the same five categories (chance level: 20%).

To binarize the S4 afferents we calculated a threshold so that each afferent is *on* for $P\%$ of the entire "training" set and *off* for the remaining $100 - P\%$ of the time. In Fig. A.8 and A.9 we show simulation results when $P = 10\%$ and $P = 30\%$. When $N = 30\%$ and more afferents (> 100), the performance of binary units becomes very similar to the one of analog units. Note that we found similar performance for $N = 30\%$ and $N = 60\%$ (not shown) and observed a drop in performance for higher values of N. Interestingly, as predicted in Section 2, the use of binary values only impact the model performance with a small number of S4 afferents (10 and 60)

We also evaluated the impact of the various approximations of the Gaussian tuning operation(see Section 2) on the model performance. We found that a *normalized dot-product* (as well as a simple *dot-product* can perform at the same level of performance as an exact *Gaussian* tuning.

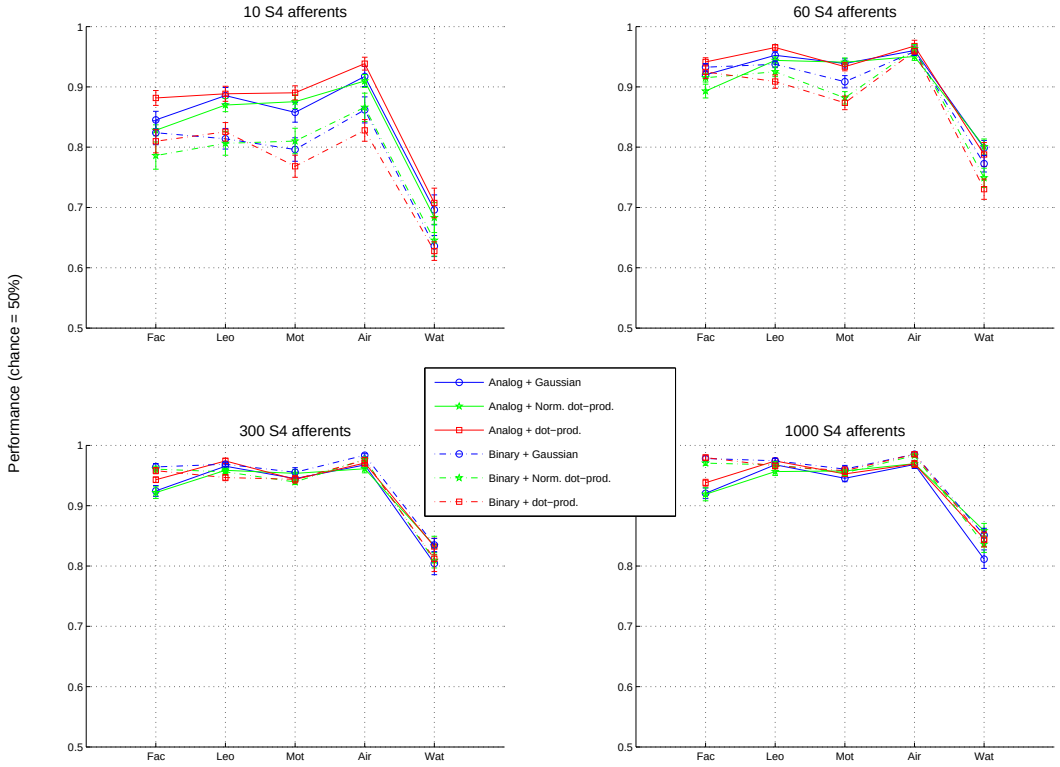


Figure A.8: Model performance on several (binary) classification problems – object present or absent (chance level 50%) – for faces, leopards, motorcycles, airplanes and watches *vs.* background images (see text). We compared the performance of several model "simplifications" such as binarizing unit responses ("on" or "off") at the top of the hierarchy and approximating the *Gaussian* tuning operation with more biophysically plausible operations such as *normalized dot-product* and simple *dot-product*, see Section 2.

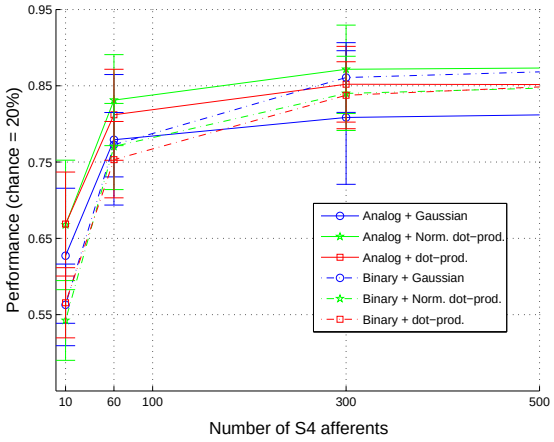


Figure A.9: Model performance on a challenging multi-class classification problem (chance level 20%) with the same five classes. As for the object present/absent task of Fig. A.8 the model performance remain fairly intact after various "simplifications".

A.6 RBF networks, normalized RBF and cortical circuits in prefrontal cortex

Instead of a standard linear classifier – which would make the architecture from IT to PFC a standard RBF network – the classifier could have a special form in order provide a normalized RBF architecture, see Fig. A.10. A standard RBF network has the form

$$f(\mathbf{x}) = \sum_i c_i K(\mathbf{x}, \mathbf{x}_i). \tag{23}$$

A normalized RBF network has the form

$$f(\mathbf{x}) = \sum_i y_i \frac{K(\mathbf{x}, \mathbf{x}_i)}{\sum_j K(\mathbf{x}, \mathbf{x}_j)}. \tag{24}$$

One of the main advantages of the latter architecture is the ability to do "one-shot" learning from IT to PFC, since the weights of the linear classifier – from the hidden units to the output – is now trivially given by the label of each example y_i instead of a set of coefficients $\mathbf{c}$ that have to be learned by minimization of an error over the set of supervised examples. The main prediction is a further normalization operation – in addition to the normalization using the inputs to IT represented by the tuning equation 1 – in IT or between IT and PFC. The circuit of Fig. 7b predicts the existence of 2 sets of VTU units, the second being exactly normalized by the activity in the first layer.

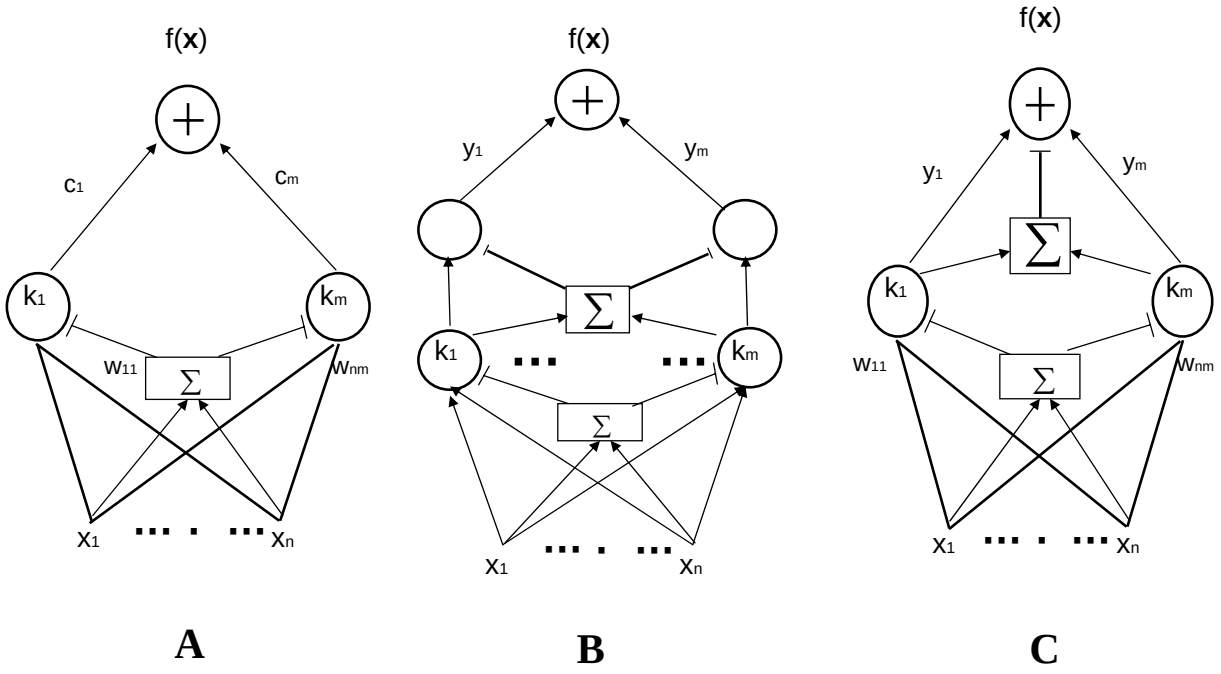


Figure A.10: A. RBF classifier with a pool cell implementing the normalization for a normalized dot product which approximates a Gaussian when further processed through a sigmoidal function; B. A normalized RBF classifier where the output normalization is done by creating normalized hidden units; C. the output normalization is done directly on the output unit.

A.7 Two Spot Reverse Correlation in V1 and C1 in the model

A.7.1 Introduction

For decades, mapping out the receptive field and finding its tuning properties of a neuron have been one of the core methods of neuroscience [Hubel and Wiesel, 1968; Ringach, 2004b]. In the study of primary visual cortex, the reverse correlation technique has proven very productive for investigating a population of linearly behaving neurons, known as simple cells. It is, however, more difficult to characterize nonlinear or higher-order behaviors, such as surround (non-classical receptive field) effects or nonlinear summations within the receptive field (for examples, [Lampl et al., 2004; Cavanaugh et al., 2002; Series et al., 2003]), and to understand the mechanisms for those effects.

In this section, we examine the sparse noise (two-spot) reverse correlation studies in V1 [Livingstone and Conway, 2003], which showed an intricate pattern of nonlinear interaction within the receptive field of direction-selective complex cells. We propose a simple normalization operation as an underlying cause for the observed nonlinear behavior, and as pointed out in [Kouh and Poggio, 2004; Yu et al., 2002; Riesenhuber and Poggio, 1999b], normalization or gain control operation is a key ingredient for generating selectivity and invariance for object recognition.

A.7.2 Two-spot reverse correlation experiment in V1

Improving upon the tradition of investigating the interactions within the receptive field of V1 cells, Livingstone and Conway [Livingstone and Conway, 2003] used two small spots of the same and different contrasts to map out the substructure of receptive field of the direction-selective complex cells. Assuming translational invariance (evidenced in [Szulborski and Palmer, 1990]), the interaction is analyzed by doing a reverse correlation over the relative distance between the two spots. Their result clearly shows the nonlinear facilitatory interactions along the preferred orientation of the cell and the suppressive interaction along the orthogonal orientation, consistent with earlier experiments [Movshon et al., 1978; Szulborski and Palmer, 1990]. Interestingly, stronger facilitation occurs when the two spots are located at some distances apart. When the two spots are closer or overlapping, there is little facilitation, making the interaction map looking like a "bowtie." See Fig. 8 in [Livingstone and Conway, 2003] for the results to compare, as well as full description of the experimental procedure and the analysis method.

The bowtie-like pattern of interaction suggests that the neural response depends on the configuration of the two-spot stimuli. Furthermore, the different level of response when the two spots are overlapping (and thus, brighter) or apart suggests that the response also depends on the luminance or contrast of the stimuli nonlinearly. As shown below, the divisive normalization of the neural response has the observed modulatory effect on the neural response.

A.7.3 Two-spot reverse correlation experiment in the model

We propose and show below that a gain control mechanism through divisive normalization of the orientation-tuned cells can account for the bowtie-like pattern of interaction within the receptive field.

We first model the simple cells with Gabor functions with orientation and phase selectivity [Ringach, 2002]. Then, we borrow and extend Hubel and Wiesel's classical model of complex cells as being combination of simple cells with the same orientation selectivity at different spatial location, giving them the translation invariance properties [Ferster and Miller, 2000]. In particular, simple cells are pooled with a maximum operation (not linear sum nor sum of squares, quadrature model), motivated by [Lampl et al., 2004; Riesenhuber and Poggio, 1999b].

As for the crucial normalization stage, we assume that the response of each complex cell is modulated by a divisive factor determined by the total response of all other complex cells of different orientation selectivity within a neighborhood, as well as itself. Similar gain control mechanisms have been proposed in various contexts before. (See [Reichardt et al., 1983; Carandini and Heeger, 1994; Carandini et al., 1997; Heeger, 1993; Cavanaugh et al., 2002; Schwartz and Simoncelli, 2001; Kouh and Poggio, 2004] and Table 6.) Such divisive normalization can be performed quite feasibly through a lateral shunting inhibition or through a shunting inhibitory by a pool cell that receives inputs from many cells in the normalization pool.

Effects	
Contrast-Dependent Responses	The cells in V1 respond to the stimuli (usually Cartesian gratings) at different contrasts in a sigmoidal fashion. At low contrast, there is almost no response, but with increasing contrast, the firing rate increases up to certain value. Beyond this saturating contrast, the firing rate changes very little. In a widely referenced study, Carandini and others [Carandini et al., 1997] have shown that such contrast-dependent responses can be nicely fitted with a divisive normalization model. Also see [Carandini and Heeger, 1994; Heeger, 1993].
Center-Surround Effect	As shown [Cavanaugh et al., 2002], the response of the neurons in primary visual cortex is modulated by normally silent "surround" or non-classical receptive field, which is considerably larger than its "center" or classical receptive field. The center-surround effects seem mostly suppressive and depend on the context. Cavanaugh and others have used a model of divisive interaction and the gain control between the smaller center and the larger surround to explain this phenomenon.
Independence in Neural Signal	Interestingly, Schwartz and Simoncelli have shown in a simulation that a similar divisive normalization scheme can achieve more independence in the neural signals, despite the dependences present in the natural image statistics [Schwartz and Simoncelli, 2001]. They have also elaborated that such divisive normalization scheme is consistent with the above two effects.
Selectivity and Invariance	As pointed out in [Yu et al., 2002; Kouh and Poggio, 2004], divisive normalization is a key ingredient of the selectivity/tuning and invariance/softmax operations needed for object recognition.

Table 6: Some effects of divisive normalization.

We applied the same two-spot reverse correlation technique as done in [Livingstone and Conway, 2003] on our model of a complex cell. For comparison, two different flavors of divisive normalization plus the case of no normalization are considered, and they are mathematically summarized as the following:[8]

$$\text{No Norm: } R_i = y_i = \max_{j \in M}\{x_{i,j}\}, \tag{25}$$

$$\text{Stimulus Norm: } R_i = \frac{y_i}{c + |S|^2}, \tag{26}$$

$$\text{Pool Norm: } R_i = \frac{y_i}{\sqrt{c + \sum_{j \in N} y_j^2}}. \tag{27}$$

$$\tag{28}$$

With divisive normalization, the response of the neuron is not just dependent on the linear summation of the independent stimuli (two spots, in this case). The response may be greater or less than the linear response depending on the magnitude of the normalization factor. The bow-tie effect suggests that the overlapping, same contrast spots produce less response than the separated spots along the preferred orientation. In terms of our model, the normalization factor will have to be smaller for an elongated configuration of the stimuli (separated two spots). Such reduction of the divisive factor can be achieved either by nonlinearly integrating the luminance of the stimuli (Eq. 26 or by adding up the responses with different orientation selectivities (Eq. 27). Because the response to the elongated configuration of the two spots will be smaller for the neuron with orthogonal orientation selectivity, whereas the response to the overlapping spots (with no particular elongation) will be similar for all the neurons, the response of the pool normalized neuron will be more facilitated for the elongated configuration of the two spots. The suppressive effects along the orthogonal direction arises from the on- and off-subregions of the simple cell kernel.

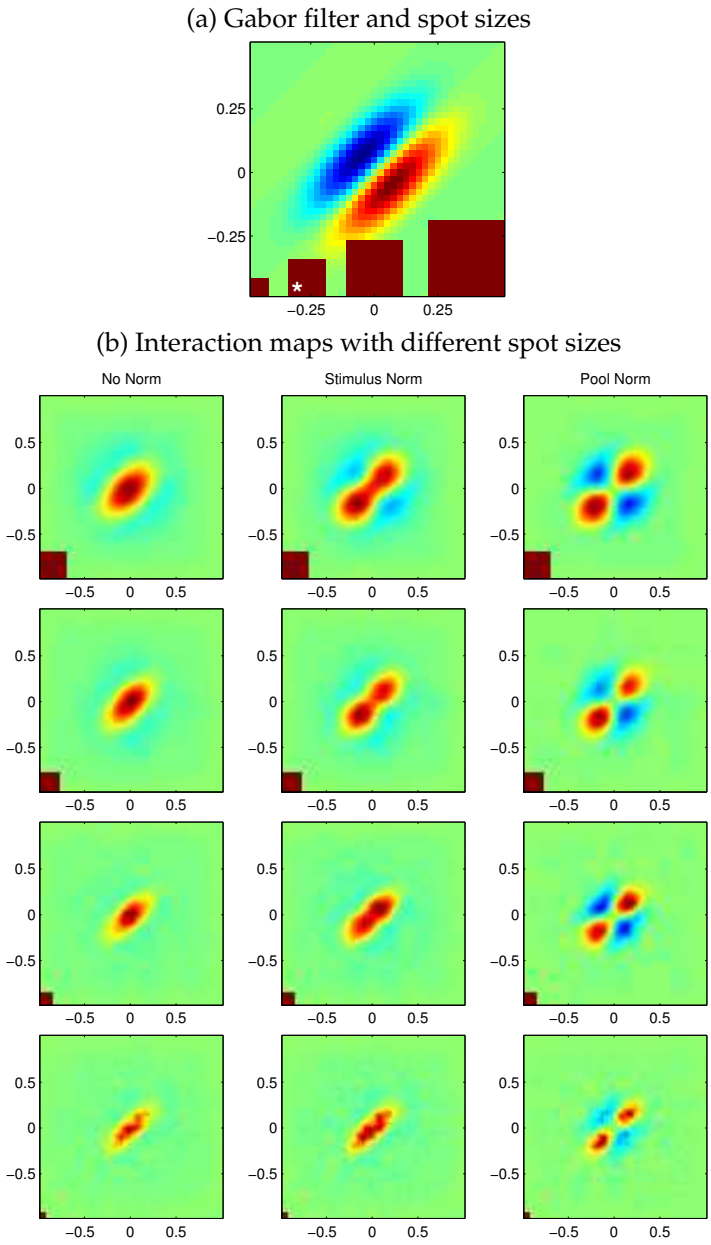


Figure A.11: (a) The receptive field structure of a simple cell is shown here, along with sample spots at four different scales as used in the simulations. (b) The effects of the divisive normalization on the two-spot reverse correlation experiments are shown here. Without any normalization (first column, Eq. 25), there is no bowtie effect. Although the stimulus normalization (second column, Eq. 26) produces separated peaks in the interaction map, the location of those peaks and the strength of the suppressive effects are not as consistent with the experimental data as the results with the pool normalization (third column, Eq. 27). The bowtie effect is readily seen with the pool normalization, regardless of the spot sizes. The size of the stimuli is shown on the lower left part of each figure. Compare with Fig. 8 in [Livingstone and Conway, 2003].

Fig. A.11b shows the simulation result of the two-spot reverse correlation method applied on the models of a complex cell with different normalizations for four different spot sizes. The bowtie-like pattern of interaction emerges from both types of divisive normalizations. However, interaction pattern from the stimulus normalization is influenced by the spot sizes, as shown in the second column of Fig. A.11b.

A simple analysis shown in Fig. A.12 illustrates that for the stimulus normalization, the peaks in the interaction map occur when two spots are just touching each other and no longer overlapping. Slightly changing the form of the stimulus normalization (for example, the power in the normalization factor) can affect the strength of the facilitatory peaks, but not the location of the peaks. On the other hand, for the pool normalization, the peak location is determined by the relative response strengths of other neurons in the normalization pool (for example, the aspect ratio of the Gabor filter can affect this), and thus, the peaks may occur when the two spots are farther apart, even with several pixels of gap.

In the experiments, the peaks occurred at the distances larger than the size of the spot themselves (M. Livingstone, personal communication), strongly supporting the pool normalization model. Furthermore, the extent and the relative strength of the suppression in the interaction maps of pool normalization are more congruent with the actual maps from V1.

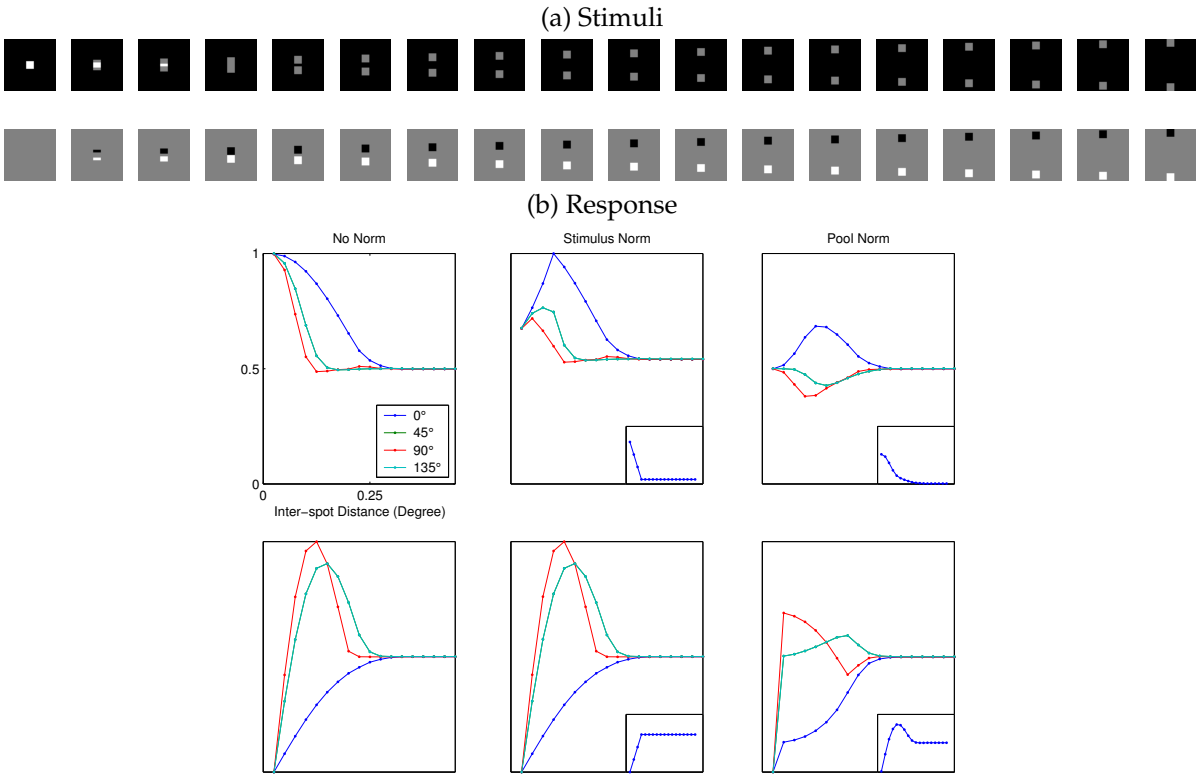


Figure A.12: (a) The two spots of the same (top) and opposite (bottom) contrasts are moved apart vertically. The size of the spots was 6 by 6 pixels, corresponding to square with an asterisk (*) in Fig. A.11 stimuli size of the stimulus (b) The responses of the model units with different normalizations (along three columns, corresponding to Eq. 25, 26 and 27) for two different stimuli conditions (top and bottom) are shown here. The insets in the second and the third columns show the normalization factor. Within each figure, model units with four different orientation selectivities are shown (because of symmetry, 45- and 135-degree orientations are overlapped). For example, in the case of same contrast stimuli (top row), the complex cell with 0-degree preference always has higher response than other cells, as expected. In the case of opposite contrast stimuli (bottom row), the cells with 90-degree preference usually have higher response, since the Gabor-like simple cells have on and off regions. *What is important here is that different normalizations produce response peaks at different inter-spot distances.* With the stimulus norm (Eq. 26), the normalization factor becomes constant when the two spots are no longer overlapping. With the pool norm (Eq. 27), the interplay of the normalization factor is more complex. Depending on how elongated the Gabor-like receptive field of simple cell is, the divisive normalization factor may stay large for a bigger range of the inter-spot distance, and hence, the response peaks appear farther away. The response peaks for opposite-contrast stimuli are also different for different normalization operations.

A.7.4 Discussion

We showed that modulating the response of a complex cell, with divisive normalization, creates a bowtie-like pattern of interaction for the two-spot reverse correlation experiments. Some types of normalization, such as stimulus normalization of Eq. 26, would depend only on the pixel-wise luminance/contrast values, whereas other types, such as pool normalization of Eq. 27 would also depend on the configuration of the stimuli, which is more consistent with the experimental data of [Livingstone and Conway, 2003].

As summarized in Table 6, there are several experimental studies of nonlinear effects in visual cortex (mostly in V1) for which the divisive normalization model was rather successfully applied. This study provided an additional supporting evidence for the normalization mechanism. Besides from the experimental evidences, there are some theoretical arguments for such gain control mechanism. In particular, the operations for selectivity and invariance, key requirements for object recognition, can be accomplished using the synaptic summation and the divisive normalization processes (last row in Table 6). Hence, it is quite feasible that similar neural circuitry for normalization may found in many different areas of cortex (not just in V1, but in V2, V4, IT, and other sensory areas). In that case, we expect to see the presence of the normalization operation through the effects like nonlinear response modulations and interactions within the receptive fields.[9]

Acknowledgment The authors would like to thank W. Freiwald, D. Tsao, and M. Livingstone for their help and useful discussions on the 2-spot reverse correlation experiments.

Notes

[8] In all the equations, x represents the response of a simple cell, y is pre-normalized response of a complex cell, and R is the final response of a complex cell. The simple cell response $x_{i,j}$ is determined by the linear operation (*e.g.,* convolution) with a Gabor summation field, as shown in Fig. A.11a. In Eq. 25, M is a set of neighboring simple cells with the same orientation selectivity i, at different positions j (*i.e.,* the response of the complex cell is determined by the max-pooling of the simple cells of the same orientation selectivity). In Eq. 26, the max-pooled response is modulated by the total luminance of the stimulus, which is represented by $|S|^2$. Such normalization may be useful for de-coupling the luminance and shape information, by dividing out the brightness of the local image patch. Finally, Eq. 27 is the normalization by the pool N of other neural responses with the same and different orientation selectivities. The small constant c avoids the division by zero. In a biological system, there equations will be true only in approximation. For example, the terms in the denominator may be raised to a different power, the effect of each neuron in the normalization pool need not be the same as in [Schwartz and Simoncelli, 2001], the responses may be subject to other types of nonlinearities, *etc.* Nonetheless, these equations are taken as exemplars from the general class of divisive normalization. Also, this mechanism can just well be implemented in the simple cell level (before the max-pooling operation for the translation invariance), as well as the complex cell level.

[9] The center-surround effects and the pattern of interaction may be caused from the same mechanism (divisive normalization), but just different in the spatial extent. The exact pattern of interaction due to normalization will be different for different areas, since the afferents are tuned differently (*e.g.,* not just Gabor filters). For example, the study of two-spot reverse correlation in V4 shows differently intricate patterns of interaction [Freiwald et al., 2005], unlike the bowtie. Hence, we note that depending on the areas to study, the probe for the normalization effect will also have to be different.

A.8 Fitting and Predicting V4 Responses

A.8.1 An Algorithm for Fitting Neural Responses

Given a V4 neuron's mean firing rate response to a set of stimuli, we would like to develop a model unit with parameters adjusted specifically to model that neuron's response. In addition, we would like the model unit to accurately predict the V4 neuron's response to stimuli the model has not seen. In this light, we can view fitting V4 neural responses as a model estimation problem. A description of the specific technique used for fitting and predicting V4 neural responses follows.

The mean firing rate response of each V4 neuron is modeled by the response of single C2 model unit by minimizing the error between the model response and the neuron's response. Four layers of the model (S1, C1, S2, and C2) have many parameters. However, the selectivity of C2 units is highly dependent on the selectivity of the S2 inputs. Therefore, the fitting procedure determines parameters at the S2 layer. S1, C1, and C2 layer parameters, such as Gabor wavelength, spatial overlap, pooling range, *etc.* are all held constant. The following properties of S2 selectivity are variable under the fitting procedure: the location and relative size of C1 subunits connected to an S2 unit, the relative weights of these subunits (w in Eq. 1), and the parameters of the sigmoid non-linearity that determines the sharpness of tuning (α, β, and its amplitude in Eq. 2).

An overview schematic of the model used in the fitting procedure is shown in Fig. A.13. A C2 unit is connected to several S2 units with identical selectivity but with receptive fields shifted over the visual field. Based on experimental findings [Pasupathy and Connor, 2001], V4 neurons maintain selectivity to translations within an area less than about .5 $\times$ the classical receptive field radius. To match these experimental findings, a C2 unit receives input from a 3×3 spatial grid of shifted S2 units. The receptive fields of adjacent S2 units overlap and are shifted by one half of the S2 receptive field radius. Each S2 unit receives input from a different, spatially translated part of the visual field. However, the selectivity parameters of each S2 unit that are pooled into the same C2 unit are identical. While there are 9 S2 units that connect to a C2 unit, only one set of S2 parameters is determined for each fit. The results of the forward fitting algorithm are shown in Fig. A.14a for V4 neuron B2 and Fig. A.14b for V4 neuron G1.

Each S2 unit is connected to a set of C1 subunits. Units in the C1 layer form a map of various sizes and receptive field positions within the receptive field of an S2 unit. C1 maps are defined with grids of three different sizes, 2×2, 3×3, and 4×4, each spanning the S2 receptive field. The sizes of individual C1 units are scaled so that within a grid, each C1 unit receptive field overlaps half of the adjacent C1 unit receptive field and the grid fills the S2 receptive field. There are a total of $2 \times 2 + 3 \times 3 + 4 \times 4 = 29$ C1 unit spatial locations with 4 oriented Gabor filters at each location for a total of 116 C1 units. S2 units are connected to a subset of the 116 (29×4) C1 units.

An analytical solution is not easily tractable due to layers of nonlinear operations in the model. However, because the model is quantitative, a numerical method can be applied to fit each V4 neuron. A forward selection algorithm is used to determine the C1-S2 connectivity while a conjugate gradient descent algorithm is used to estimate the weights and the sigmoid parameters of the S2 selectivity function. Therefore, during each iteration of the forward selection algorithm, the combination of n subunits with the lowest mean squared error between the V4 neuron's response and the C2 unit's response is selected. In the next iteration step, every possible configuration with $n+1$ C1 subunits (the winning configuration from the previous iteration plus an additional C1 subunit) is examined to find a better fit.

Predicting Within Class Stimuli A cross-validation methodology is used to predict a neuron's response to within-class stimuli. The set of measurements for a stimulus class is divided into n randomly selected and equal sized folds. The forward selection algorithm is used on fit a model unit on $n - 1$ of the folds (the training set) and the resulting model is used to predict the V4 neuron's response to the remaining fold (the test set). The response for each V4 neuron measured with the boundary conformation stimulus set, B1-B4 [Pasupathy and Connor, 2001], is analyzed with a 6-fold cross-validation methodology (305 points for training, 61 for testing). For V4 neurons measured with grating stimuli, G1-G4 [Freiwald et al., 2005] a 4-fold cross-validation is used (240 points for training, 80 for testing). For each split of the data, the forward selection algorithm terminates, or stops adding C1 subunits, when the mean-squared-error on the training set decreases by less that 1%. The predictions on each of the data folds are recombined to form a set of predictions for the entire stimulus set.

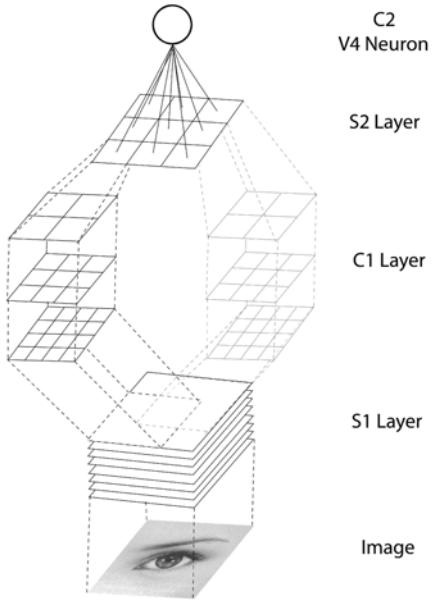


Figure A.13: Schematic diagram of the model V4 neuron. The response of the C2 unit in the top layer is comparable to measured V4 neuron responses. Units in the 'C' layers perform the maximum operation on their inputs, while units in the 'S' layers compute a Gaussian-like template matching function (normalized dot product with sigmoid transfer function). The C2 unit is connected to a 3×3 spatial grid of S2 units with identical tuning properties. S2 units are connected to a set of complex V1-like C1 units that are determined by the fitting procedure. C1 units form grids of 3 different sizes: 2×2, 3×3, and 4×4. C1 units compute the maximum response of simple V1-like S1 units with slightly translated and scaled receptive fields. The C2 unit's receptive field is the entire input image.

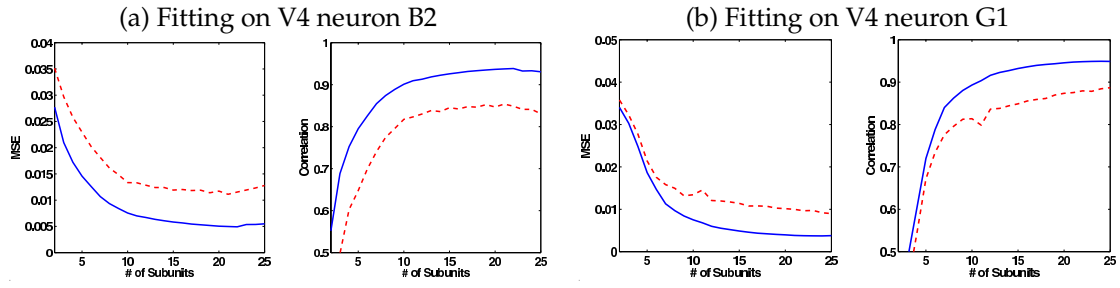


Figure A.14: (a) Evolution of the forward fitting algorithm for V4 neuron B2. The x-axis in both plots indicates the number of C1 units connected to an S2 unit for that iteration of the algorithm. The left plot graphs the mean-squared-error as a function of the number of C1 subunits. The solid blue line indicates the average mean-square-error (mse) on the training set for each fold. The dashed red line indicates the average mse on the test set for each fold. For each fold the model is trained on 5 parts of the data (train set) and used to predict the remaining part (test set). The corresponding V4 neuron is shown in Figure 4 of [Pasupathy and Connor, 2001]. (b) Evolution of the fitting algorithm for V4 neuron G1 measured with the grating stimulus set. The plotting conventions are the same as in (a), except a 4-fold cross-validation was used for this data set.

Fig. A.15 shows plots of the model's predicted response plotted against the V4 neuron's measured response. Fig. A.15a displays the prediction for the same neuron, B2, as in Fig. A.14a, and Fig. A.15 displays the prediction for the same neuron, G1, as in Fig. A.14b. A perfect prediction would result in points falling along the diagonal (indicated by the green line). The predicted values and measured values have been scaled so that the training set is between 0 and 1 (this produces some measured responses greater than 1 and less than 0; note that the fitting algorithm can, in theory, predict responses that are higher than the values included in the training set).

Predicting 2-Spot Interactions To predict V4 responses to novel stimulus classes, a model unit is fit to a V4 neuron's response to one set of stimuli and the resulting model unit is then used to predict the V4 neuron's response to a novel stimulus class. In this case, the number of subunits for each model unit is determined by the average number of subunits found on the cross-validation method over all neurons (see Table 4). As a result, a model unit is fit using all data points for a stimulus set, in this case the grating stimulus set, and the forward fitting algorithm is halted when 18 subunits are found. The resulting model unit is used to predict the response to a novel stimulus class, in this case the 2-spot stimuli. Neurons G1, G2, G3, and G4 are analyzed with this procedure. A model is found for each neuron using responses on the grating stimulus set and the model unit is used to predict the 2-spot reverse correlation map.

The 2-spot interaction maps are stable for a range of model parameters. As the forward fitting algorithm adds subunits to the model, the predicted 2-spot interaction map can change. Fig. A.16 shows the predicted interaction maps for V4 neuron G1 as additional subunits are added to the model. The top-left panel shows the prediction for 2 subunits and additional subunit predictions are plotted from left to right and from top to bottom; the bottom-right panel shows the prediction for a model with 25 subunits. Interestingly, the predictions for models containing between 15 and 20 subunits all resemble the measured 2-spot interaction map for this neuron. This indicates that the prediction is stable over a range of subunits. The evolutions for both G2 and G4 show similar stability in the prediction.

Can S2 units model V4 responses? Thus far we have used C2 units to model V4 responses; however, could S2 units produce the selectivity and invariance exhibited by V4 neurons? To test this hypothesis a one layer network, on top of the C1 unit map, was used to fit a stimulus set derived from the measurements of V4 neuron B3. The stimulus set was designed to include strong constraints on both selectivity and translation invariance. The response of V4 neuron B3 to the main boundary conformation stimulus set was augmented with simulated responses to translated stimuli selected from the main set. These stimuli were chosen to span the full range of firing rate observed in the neuron and were simulated to maintain selectivity over translation (the response to the centered stimulus was reproduced over a grid matching translation invariance observed for V4 neuron B3). Ten stimuli over a 5×5 translation grid (see Figure 4.4 for an example stimulus) were added to create a total of 606 stimuli (selectivity + invariance) for the experiment. The one layer network consisted of a linear support vector machine (SVM) for regression. The inputs to the SVM included all C1 units with a 2×2 grid feeding into the S2 layer (a total of $3\times3\times4\times4 = 144$ features). The results of the SVM network and the model C2 unit are summarized in Table 7. The correlation for the SVM fit on this data set with the simulated neuron response was 0.60. A model C2 unit fit on the same stimulus set and C1 population, achieved a correlation of 0.79. As a baseline, the SVM and the C2 unit were also fit on a stimulus set without the translated stimuli (selectivity only). Both the SVM and the C2 unit are able to produce high correlations on the selectivity set. However, the SVM network is unable to maintain this level of correlation over the stimuli that test for invariance. Therefore, for models of this complexity level, a two layer network, such as an S2 layer followed by a C2 unit, is necessary for achieving the selectivity and invariance exhibited by V4 neurons.

A.8.2 What Mechanisms Produce 2-spot Interaction Maps?

The model can also provide insight into the underlying mechanism of 2-spot interaction maps. We hypothesize that 2-spot interaction maps exhibited by V4 are a result of relative position coding and not the orientations of subunit afferents. (Note that as pointed out in Appendix A.7, the 2-spot interaction maps of complex cells in V1 are due to their orientation selectivities, in contrast to the area V4.) In order to test this hypothesis, the same procedure for predicting 2-spot interaction maps was used on V4 neuron G1 and

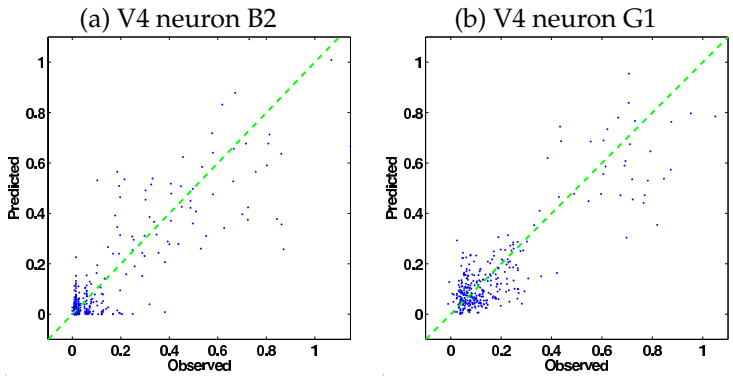


Figure A.15: (a) Model predicted response plotted against measured V4 neuron B2 response for within-class predictions (over the boundary conformation stimulus set). Each point is plotted at the actual measured response of the V4 neuron to a stimulus and at the model predicted response for the same stimulus. Perfect prediction would result in points along the diagonal, indicated by the green line. The V4 neuron for this figure is the same as in Fig. A.14a and has been adapted from Fig. 4 in [Pasupathy and Connor, 2001]. (b) Model response plotted against measured V4 neuron G1 response for the predicted points (over the grating stimulus set). V4 data is courtesy of Freiwald, Tsao, and Livingstone [Freiwald et al., 2005].

Stimulus Set	SVM	C2 Unit
Selectivity + Invariance	0.60	0.79
Selectivity Only	0.83	0.82

Table 7: Comparison of a linear SVM regression network to the C2 unit model. Correlation coefficients are shown for each model on two stimulus sets. One that contains both the main boundary conformation stimulus set and a large translated stimulus set (selectivity + invariance), and a stimulus set that contains only the main boundary conformation set (selectivity only). While both networks achieve high correlations on the selectivity only stimulus set, only the C2 unit is able to achieve a high correlation over the stimulus set that tests for both selectivity and invariance. This result indicates that for models of this complexity, a S2 layer followed by a C2 unit is necessary for achieving both the selectivity and invariance measured in V4 neurons.

two perturbations of the resulting model were tested. This results in a model that accurately reproduces both the grating response, with a correlation of 0.93, and the 2-spot interaction map of V4 neuron B1. The S2 subunit schematic and the resulting 2-spot interaction map are shown in the first row of Fig. A.17. In the first perturbation, shown in the second row of Fig. A.17, the orientation of each subunit is rotated by $90°$. The resulting model produces a 2-spot interaction map that is nearly identical to the original model. However, this model now has a very low correlation with the grating response: 0.10 between the response of V4 neuron G1 and the model unit over the grating stimulus set. In the second perturbation (the third row of Fig. A.17) the subunits of the original model are repositioned while preserving the subunit's orientation selectivity. For example, a unit in the upper right in the original model is moved to a position in the lower right (positions were chosen randomly). This perturbation results in a markedly different 2-spot interaction map and also produces a model with a low correlation to the V4 neuron's grating response: 0.14. Therefore, it appears that in the model, 2-spot interaction maps result from subunit spacing, or the relative position of subunit afferents.

A.8.3 A Common Connectivity Pattern in V4

According to the reports of Pasupathy and Connor [Pasupathy and Connor, 2001], many V4 neurons exhibit tuning to high curvature boundary elements, and a large portion of those neurons have strong relative position tuning in the boundary conformation stimulus set (see Fig. 4.2.2). Therefore, it is interesting to examine the substructure of model units that display such tuning. For this purpose, the response of V4 neuron B3 to the relative position boundary conformation stimulus set was perturbed (with Gaussian white

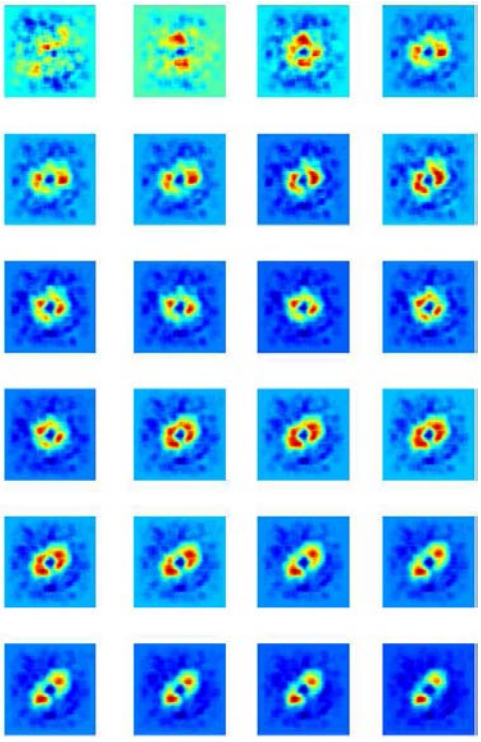


Figure A.16: Panels display the predicted 2-spot interaction map of V4 neuron G1 for each iteration of the fitting algorithm. A model unit is fit to the grating stimulus response of the V4 neuron and the 2-spot interaction map is generated for each additional C1 subunit. Models with additional subunits are plotted from left to right and from top to bottom starting for a 2 subunit model and ending with 25 subunit model. The predictions for models with 15 to 20 subunits show qualitative agreement with the experimental 2-spot interaction map.

noise) to generate additional responses, 30 in total, that exhibit the same tuning pattern. Model units were then fit to the response patterns of the 30 generated neural responses. The weights, w, or subunit centers were clustered using a k-means algorithm to determine 5 groups of models.

Fig. A.8.3 shows the results of the clustering algorithm. The number of models clustered into each of the 5 centers is shown in a histogram in the lower right. A similar activation pattern is present in each of the 5 centers: a horizontal subunit in the upper right and -45° diagonal unit in the lower left. These activation patterns represent models that would exhibit the same distinctive tuning pattern of many neurons found in V4. Therefore, we expect that some connectivity patterns (in terms of similar relative positions and orientations of subunits) are prevalent in V4.

Acknowledgment The authors would like to thank our collaborators: A. Pasupathy and C. E. Connor (on boundary conformation experiments); and W. Freiwald, D. Tsao, and M. Livingstone (on the grating and 2-spot experiments).

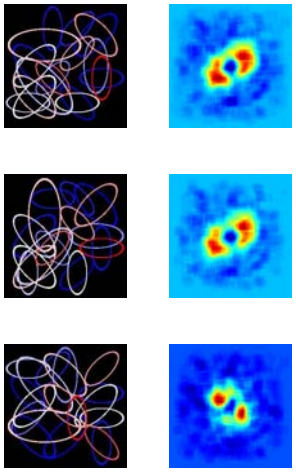


Figure A.17: Results of model perturbations that indicate relative position coding as a cause of 2-spot interaction maps. The first row displays the subunit schematic and the 2-spot interaction map for a model that shows high correspondence with V4 neuron G1 over both the grating stimulus set (correlation = 0.93) and the 2-spot interaction map. The second row shows the subunits and 2-spot interaction map for a model identical to the model in the first row, but with the orientation selectivity of each subunit rotated by 90 degrees. The 2-spot interaction map for this model closely resembles that of the first model, yet shows a poor fit with the grating response (correlation = 0.10). The last row shows a model identical to the first, but with the position of each subunit randomly shifted. The resulting model shows a drastically different 2-spot interaction map and a poor correlation with the grating response (correlation = 0.14). These results indicate that within the model, 2-spot interaction maps are a product of the relative position of subunit afferents, and not of the orientation selectivities of the subunit afferents.

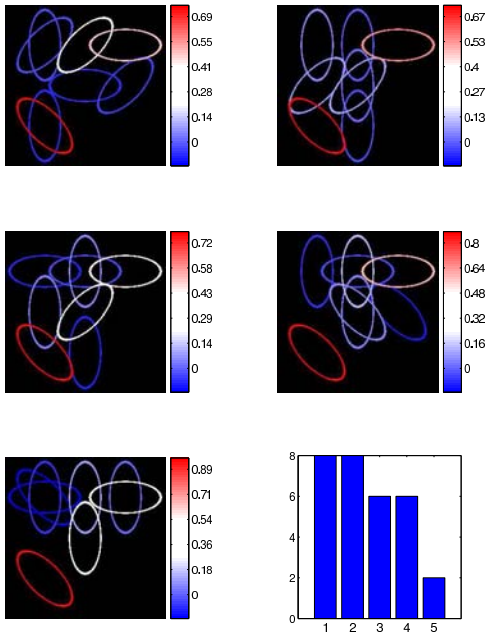


Figure A.18: Model clusters showing S2 weights that exhibit the relative position tuning described in [Pasupathy and Connor, 2001]. Five cluster centers are shown and the histogram of labels is shown in the lower right corner. These activation patterns represent models that exhibit a particular tuning pattern found in V4. Similar connectivity may be prevalent in area V4.

A.9 Fast readout of object information from different layers of the model and from IT neurons

In this appendix we provide further information and support for our comparison of the decoding performance using neuronal responses versus model responses. In particular, here we show that the later stages of the model can well account for the properties of IT neurons observed in electrophysiological recordings from macaque monkeys.

A.9.1 Methods

Experimental recordings in IT neurons We compare the results to an experiment in which multi-unit spiking activity and local field potentials were measured in the inferior temporal cortex of macaque monkeys [Hung et al., 2005a]. Further details about the observations in IT can be found at `http://ramonycajal.mit.edu/kreiman/resources/ultrafast/`. Recordings were made from two monkeys (*Macaca mulatta*). We used a set of 77 complex objects rendered in pseudo-random order in grays scale A.19. Objects were divided prior to the recordings into 8 classes: toys, foodstuffs, human faces, monkey hand/body parts, monkey faces, vehicles, white boxes, and synthetic images of cats and dogs. Object images (3.4 deg) were presented during passive fixation. Object images were presented on the screen for 93 ms followed by a 93 ms blank period. Object images were not normalized for mean gray level, contrast or other basic image properties. It is possible to partially read out object category based on some of these simple image properties. Penetrations were guided by structural MR images of each monkey and made over a 10x10 mm area of the ventral surface surrounding AMTS (Horsley-Clark AP: 10-20 mm, ML: 14-24 mm) using glass-coated Pt/Ir electrodes (0.5-1.5 MΩ at 1 kHz). Spiking activity (400 Hz-6 kHz) and LFPs (1-300 Hz) were amplified, filtered, and stored.

Figure A.19 shows a sample of the images and scale/position transformations that were used in this study.

Classifier analysis For any given recording site s ($s = 1, ..., N$); let r_{ij}^s denote the response during repetition i of object image j, (typically $i = 1, ..., n_{rep}$, $n_{rep} = 10$, $j = 1, ..., 77$; for the study of invariance to scale and position $j = 1, ..., 385$). In contrast to the simulations, the number of sites available for decoding is a major limitation, particularly when using single electrode recordings as in the present case. The number of sites was $N = 364$ for MUA data, $N = 71$ for the invariance study with MUA data, $N = 315$ for LFP data, $N = 45$ sites for the invariance study with LFP data.

An important question to understand neural coding is to decipher what are the temporal scales over which neurons can convey information. The current implementation of the model does not include any considerations about dynamics (with the exception of the biophysical models of Section 5). Thus, we do not discuss the dynamics of the responses (or the separation of spikes into bursts and isolated spikes, or the determination of the stimulus onset time from the population activity) in any detail here except to say that we observed that information conveyed in single bins of 12.5 ms duration can convey accurate, reliable and selective information for categorization and identification. Thus, in our case, for each unit, we have a scalar representing its response to each image (this scalar turns into a vector when considering the dynamics).

When considering a population of multiple units we simply concatenated the responses from the individual units. It should be noted that this concatenation step assumes independence among different neurons, an assumption that needs to be revisited upon availability of simultaneous recordings from multiple neurons. It is quite possible that correlations between neurons - which we cannot detect with our independent recordings - may contain additional information and reveal additional aspects of the neural codes [Steinmetz et al., 2000] (see however, [Aggelopoulos et al., 2005]). Our estimates are thus likely to constitute a lower bound on the information represented by small populations of neurons in IT. It is interesting, however, that even under these conditions we obtain such a high degree of accuracy in decoding visual information. Again, this issue is not considered in the current version of the theory.

The resulting input was used as input to the decoding classifier (see below). The dimensionality of the input is therefore given by the number of units (N) in the current case. We used $N = 1, 2, 4, ..., 256$ sites (e.g. Figure 4.14. Regularization becomes critical, particularly for high dimensions.

The data were always divided into a training set and a test set. The separation of the training and test sets depended on the particular question and experiment. In the IT recordings, neurons show apparent

Figure A.19: Sample of the images and scale/position transformations that were used to decode the information contained in the model responses and compare the results against the ones observed from IT recordings.

trial-to-trial variability. We therefore started by asking whether the population activity could generalize over this presumed variability by setting the training set to comprise 70 % of the 10 repetitions of each object image while the test set included the remaining 30 %. In the case of studying the extrapolation to different pictures within a class (Figure A.23), training was performed on 70 % of the pictures and testing on the remaining 30 % of the pictures.

We focused particularly on two tasks which are usually considered different (even if the difference is mostly semantic and they are completely equivalent from a computational point of view): *classification* and *identification*. For classification, the picture labels indicated which out of 8 possible classes the picture belonged to (toys, foodstuffs, human faces, monkey faces, hand/body, vehicles, white boxes, cats/dogs). Chance was therefore $1/8$. For identification, the picture labels directly indicated the identity of the image (77 possible labels). Chance was therefore $1/77$. We used a one-versus-all classification scheme. We trained one binary classifier for each class against the rest of the stimuli. For a novel input, the prediction was given by the classifier with the maximum output. Classification performance, as shown on all the plots, indicates the proportion of correct decoding for test data (i.e. data not seen by the classifier during training). We also compared the classification results against those obtained by arbitrarily assigning pictures to groups in a random fashion.

We compared the performance of different statistical classifiers including Fisher linear discriminant (FLD) classifier, Support Vector Machine (SVM) using linear or Gaussian kernels and Regularized least squares classifier (RLSC, which is a linear classifier). The SVM classifiers yielded the best performance. Throughout the main text, we used the SVM with linear kernel because its architecture can be easily implemented in cortex as a thresholded sum of weighted synaptic inputs.

In most of the graphs described throughout the text the sites used as input to the classifier were randomly chosen from the total set of N sites. This random choice of the sites was repeated at least 20 times

(and in most cases 50 times) and we report the average obtained from all of these random neuronal subpopulations. As a very simple approach to feature selection, we also considered the situation where sites were chosen if they were particularly "good" for the classification task. For this purpose, we defined the signal to noise ratio (SNR) for a site s ($s = 1, ..., N$) and a particular stimulus group g ($g = 1, ..., G$). Sites were ranked for each group based on their SNR values. To select n "good" sites, we iteratively chose the one with the highest SNR, then a different site with the highest SNR for a different group and so on.

In most models of visual object recognition and in many computer vision algorithms, the ability to generalize across (e.g.) changes in scale and position may be 'learned' from visual experience of a specific object (see Section 1). To examine this possibility, we tested the ability of individual IT neurons to generalize their selectivity across the same changes in position and scale for novel objects that were previously unseen by the animal. We directly compared the degree of position and scale invariance for novel objects (never viewed by the animal until the time of recording, 10 novel objects per recording site) vs. familiar objects (each previously viewed by the animal 500 times at each of the transformed scales and positions and 30,000 times at the standard scale and position). This comparison could not be made at the population level in the IT recordings because new objects had to be used for each single-electrode recording session. Instead, we made a direct comparison of position and scale invariance of novel versus familiar objects at each site. We compared the generalization across position and scale with novel objects (10 novel objects per recording site, 13 sites and 130 objects total) with that observed for familiar objects. We observed that individual IT neurons could well generalize over scales and positions (at least for the scales and position transformations tested here) to novel objects. It remains to be determined whether such invariance derives from a lifetime of previous experience with other related objects (that is, scale and position invariance would be learned) or from innate properties of the visual system (that is, scale and position invariance could be hardwired) or some combination of both. In any case, the observation that the adult primate IT population has significant position and scale invariance for arbitrary 'novel' objects provides a strong constraint for any explanation of the computational architecture and function of the ventral visual stream. This observation is well accounted for by the theory described in this manuscript.

We studied the robustness of the code to neuronal drop-out, important biological source of noise. Neuronal drop-out is meant to simulate neuronal or synaptic death. To study the robustness to neuronal drop-out, we trained the classifier as explained above. During testing, a given percentage of the sites were randomly removed and the classifier was forced to reach a decision using the remaining sites.

Free parameters When using computational models, it is generally important to describe the number of free parameters that are changed to attempt to fit the observations. The number of free parameters can have large impact in the accuracy of the fit but it can also affect generalization if not controlled carefully. The design and parameters of the model have been discussed elsewhere in this memo. For all the figures shown in this section, we did not change *any* parameters in the model. Thus, there are no free parameters in the model. The goal was to "record" from the model as an investigator would record from neurons. Training of the classifiers had multiple free parameters (see above) but this step was identical for electrophysiological recordings and model recordings.

A.9.2 Further observations

Scale and position invariance for single units In Section 4.3, we showed that the population response showed invariance to scale and position transformations. Here we illustrate the progressive increase in the degree of invariance for *individual* units from S1 to C2 by comparing the responses of different layers of the model to the same 77 objects used in the IT read-out experiment shown at different scales and positions (Figure A.20). For direct comparison, we also show in the same format the results obtained from spiking (MUA) and local field potentials (LFP) recordings in IT cortex [Kreiman et al., In press]. The LFP recordings are close to the model units somewhere between S2 and C2, consistent with the idea that these units represent the input to IT cortex. The MUA recordings show weaker invariance than the C2 units in the model according to this measure. This is likely to be, at least partly, due to the very large receptive fields of C2 units.

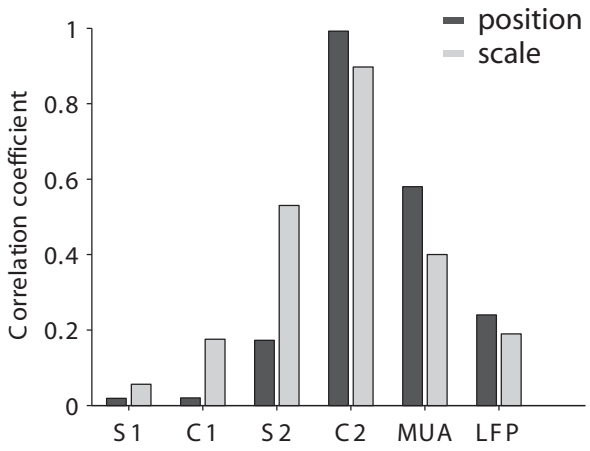


Figure A.20: Comparison of the responses to the 77 objects used in the read-out task across scales (gray) or across positions (black). The results shown here correspond to the average over 10,000 S1 units, 10,000 C1 units, 10,000 S2 units, 256 C2 units and 364 IT recording sites for multi-unit activity (MUA) and local field potentials (LFP). For each unit, we computed the Pearson correlation coefficient between the average response to each of the 77 objects at the center of gaze and standard scale versus the corresponding responses when the objects were scaled or translated. For scale invariance, the results illustrated here show the average over two scales (0.5 and 2 times the standard scale); for position invariance the results show the average over two positions (2 and 4 degree shifts from the center of gaze).

Scale and position invariance for the population We showed in the main text the degree of extrapolation to scale and position changes for the C2 units in the model 4.15. Here we show the results for a population of units at different stages of the model A.21.

Similarity between pictures based on unit responses The groups used in the "categorization" task were defined based on semantic categories before the experiments. However, comparison of the units' activity across pictures showed that the responses to two pictures within the same group were in general more similar than the responses to two pictures in separate groups. Consequently, comparison of the population response vectors for pairs of pictures yields clear delimitation of some of the same categories used in the classification experiments (Figure A.22. Unsupervised clustering of the C2 unit responses using k-means clustering based on the correlation coefficients shown in Figure A.22 yields groups that are very close to the ones used in the read-out task. Thus, the C2 units population yields similar responses to objects that are similar.

Extrapolation to novel pictures within a category This mapping between the stimulus domain and the unit response domain suggests that it would be possible to extrapolate to decode novel pictures within the same category. Indeed, we also obtained high performance levels when we trained the classifier with a fraction of the pictures (70 %) and tested the ability of the classifier to predict the category of the remaining 30 % of the pictures (Figure A.23).

A.9.3 Predictions

One of the features of working with a computational model is the possibility of making quantitative predictions that can be tested in the lab. The goal of computational predictions is not to eliminate empirical testing but to suggest concrete experiments that can be tested to verify or refute the predictions. In some

cases, (e.g. electrophysiological experiments in behaving primates), experiments may require substantial amounts of time or may be difficult to perform. Computational predictions (if they are good) can provide hints to suggest which experiments are most relevant to focus on.

In the previous section, we showed that the model can account for many of the empirical observations observed from recordings in macaque IT cortex. Here we show some related predictions that go beyond the empirical observations and in turn suggest novel experiments.

Extrapolation to many images In the extreme of using only a very small number of examples to evaluate recognition, high performance could be obtained due to overtraining or focusing on anecdotal aspects of each object. Therefore, it is important to show that a recognition system can generalize to large data sets. Our own visual system seems to have a very large capacity to represent large numbers of objects and object categories. Most of the results shown in Section readout.sec were based on a set of 77 objects divided into 8 possible categories.

In Figure 4.16 we tested the categorization performance of the model using 787 objects divided into 20 possible object categories. The procedure was the same as the one described above. Basically, for each object category, we trained the classifier on the responses of C2 units to 10 % of the objects from each category presented at random positions in the image. Classification performance was then evaluated on the remaining 90 % of the images also presented at random positions. Thus, the results shown in Figure 4.16 include extrapolation to larger object sets, extrapolation to novel objects within known categories (see also Figure A.23) and robustness to position changes.

Also, a study using 4075 C2 units on a very large set of 101 categories (with 40 to 800 objects per class) recently obtained a performance of 42 % for multi-class [Serre et al., 2005c].

These observations suggest that, at least in principle, the architecture in the current model is capable of supporting recognition using large numbers of objects and object classes.

Robustness to background changes Another important aspect of natural vision is that objects are generally embedded in natural changing backgrounds. We generally do not have trouble discriminating salient objects even if they appear in very different backgrounds (e.g. a person can be recognized from his face in the library, in the park or at work). An exception to this claim may be those cases where the object is particularly similar to the background (and in the absence of movement). This is the principle of camouflage.

We explored read-out of object category or identity after embedding the objects in multiple complex backgrounds. Here we did not particular make any efforts to ensure that the resulting images "made sense" to a human observer; thus, a car could be floating above a building. It may be important to actually control this aspect much more careful in experiments with monkeys or humans. Also, the objects were generally clearly distinguishable from the background and we did not particularly try to camouflage the objects.

At least to a first approximation, we observed that we could read out the identity and category of the objects even when they were embedded in a very different background images (Figure 4.17).

Extrapolation to multiple objects We also extended the observations to the case where the images contain more than one single isolated object. We tested the performance of C2 units in the model to perform object recognition in the presence of two or three objects. For each of the 77 objects, we trained the classifier with images containing that object and other objects (at least 8 examples with other objects and 5 different relative positions of the two objects). Performance was then evaluated using novel object combinations that were not used for training. We exhaustively used all object pairs or triplets (giving a total of $\binom{77}{2}$ examples for object pairs and $\binom{77}{3}$ examples for object triplets). Figure A.24 **A** and A.25 **A** show examples of the object pairs and triplets used.

We first asked whether we could reliably identify or categorize a single object even in the presence of other objects in the image. For this purpose, we considered a single prediction from the classifier and a hit was defined by a match from the classifier's prediction to *any* of the objects present in the image. This analysis showed that we can robustly perform object recognition with the activity of a population of C2 units even in the presence of multiple objects (Figure A.24 **B** and Figure A.25 **B**).

We next considered whether we could describe the multiple objects present in the image by considering more than one prediction from the classifier. In the multi-class approach, we considered only the best

prediction across all binary classifiers. Here we considered the two best predictions (Figure A.24 **C-D** and Figure A.25 **C-D**) or the best three predictions (Figure A.25 **E-G**). We also asked the classifier to estimate the number of objects present in the image. This could probably be done by directly training a classifier with the number of objects as a label. With the current images, the absence of a background implies that this task could be easily accomplished by considering the overall contrast or th intensity of the images. Therefore, we used a different approach where we considered the number of active binary classifiers. The total number of binary classifiers that were activated converged to the actual number of objects for 256 units.

Describing all the objects in an image in a brief glimpse (in the absence of multiple fixations), can be a difficult task, even for our recognition system [van Rullen and Koch, 2003a]. Here we show that the performance of a small population of C2 units is highly above chance for describing multiple objects.

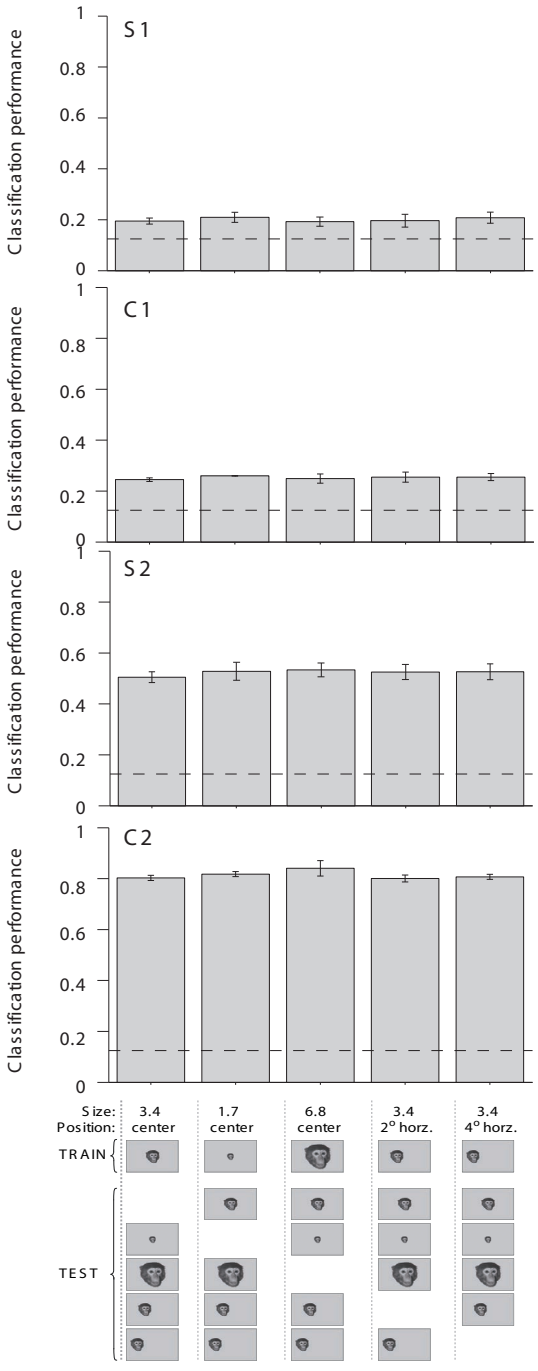


Figure A.21: Classification performance upon training and testing with the same images shown at different scales and positions for different layers of the model (from S1 through C2). The first column shows the performance obtained upon training the classifier with the 77 images at the center of gaze and 3.4 degrees size and testing on the small or large scale and shifted versions. The second column shows the performance obtained upon training the classifier using the small scaled (1.7 degrees), etc. The horizontal dashed line indicates chance performance level.

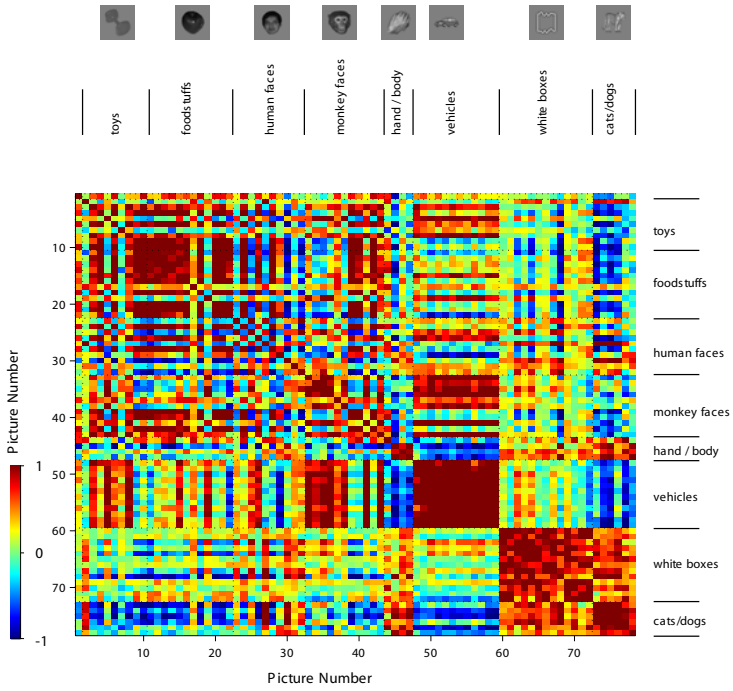


Figure A.22: Similarity between pictures based on the C2 unit activity. Each entry (i, j) in the matrix shows the Pearson correlation coefficient between the vectors containing responses of all C2 units to picture i and picture j. The scale of the correlation coefficients is color-coded (see scale at the bottom of the figure). The diagonal is 1 by definition. The dimensions of the matrix are 78 x 78 (picture 1 is a blank image). The dashed lines divide the groups of pictures used in the classification task (see text for details). A representative example from each group is shown above the matrix.

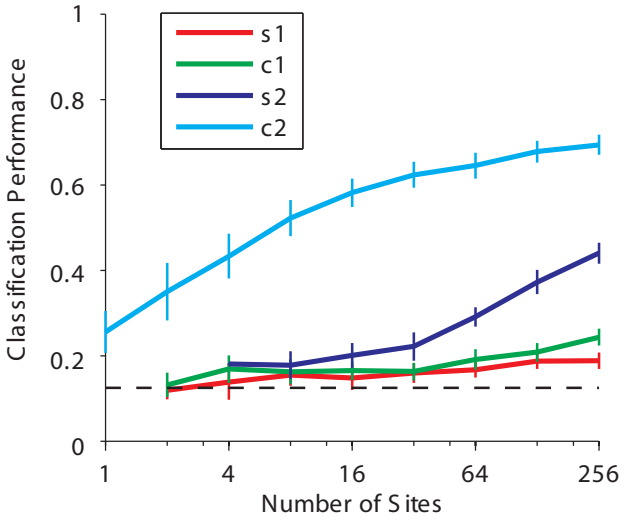


Figure A.23: Classification performance for extrapolating object category for novel pictures (never seen by the classifier during training) within the same categories used for training. The format follows the same as in Figure 4.14.

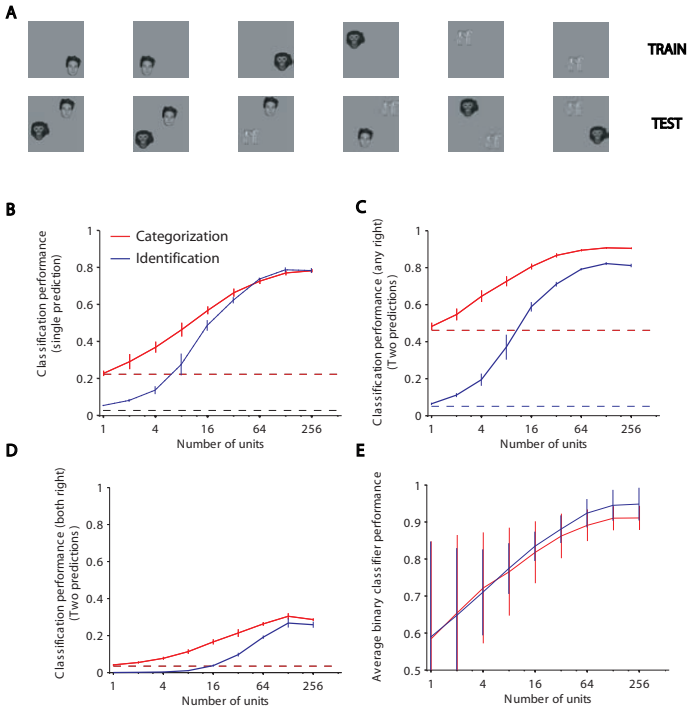


Figure A.24: Classification performance in images containing two objects. **A** Examples of the images containing multiple objects. In the top row here we show that single isolated objects were used to train the classifier. However, as discussed in the main text (see Section 4.3.3), we could also achieve similar results upon training the classifier with images containing multiple objects. **B** Classification performance using the best prediction of the classifier for categorization (red) and identification (blue). A match to *any* of the two objects present in the test image was considered to be a hit. **C** Here we considered the two best predictions of the classifier. A match from any of these two predictions to any of the two objects present in the image was considered to be a hit. **D** Using two predictions as in **C**, here a match was defined as being able to correctly describe the two objects present in the image. The chance level for identification is very low (roughly $\frac{2}{77*76}$). **E** Average performance of each of the binary classifiers (8 for categorization, red, and 77 for identification, blue).

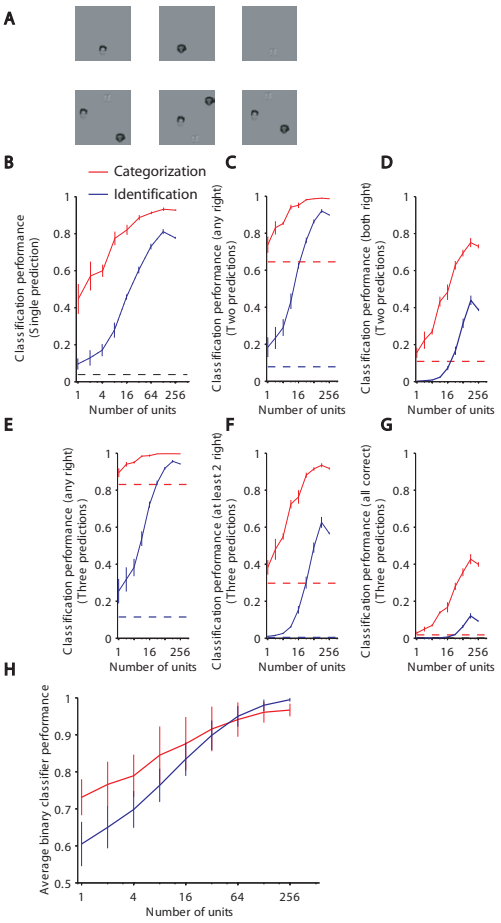


Figure A.25: Classification performance in images containing two objects. **A** Examples of the images containing multiple objects. In the top row here we show that single isolated objects were used to train the classifier. However, as discussed in the main text (see Section 4.3.3), we could also achieve similar results upon training the classifier with images containing multiple objects. **B** Classification performance using the best prediction of the classifier for categorization (red) and identification (blue). A match to *any* of the three objects present in the test image was considered to be a hit. **C** Here we considered the two best predictions of the classifier. A match from any of these two predictions to any of the three objects present in the image was considered to be a hit. **D** Using two predictions as in **C**, here a match was defined as being able to correctly describe (red, categorization; blue, identification) two of the three objects present in the image. **E** Here we considered the three best predictions of the classifier. A match from any of these three predictions to any of the three objects present in the image was considered to be a hit. Note that the chance levels are very high, particularly for categorization (red). **F** Using three classifier predictions as in **E**, here a match was defined as being able to correctly describe (red, categorization; blue, identification) two of the three objects present in the image. **G** Using three classifier predictions, here a match was defined as being able to correctly describe all three objects present in the image. The chance levels for identification are very low. **H** Average performance of each of the binary classifiers (8 for categorization, red; 77 for identification, blue).

A.10 Categorization in IT and PFC

One of the model predictions is that all visual areas up to IT are trained independent of a particular task, forming a general representation of the world, and only higher areas such as PFC encode task-specific information on the basis of the general representation in IT. To test this prediction, Freedman and Miller performed physiology experiments providing experimental population data for both PFC and IT of a monkey trained on a "cat/dog" categorization task [Freedman et al., 2001]. Using the same stimuli as in the experiment, we analyzed the properties of view-tuned units in our model, trained without any explicit category information, and compare them to the tuning properties of experimental IT and PFC neurons. This is work done in collaboration with Maximilian Riesenhuber and David Freedman and has been published in a more extended version as an AI and CBCL memo [Knoblich et al., 2002].

Category tuning We use three measures to characterize the category-related behavior of experimental neurons and model units: the between-within index (BWI), the class coverage index (CCI) and the receiver operating characteristics (ROC).

In particular, we analyzed a total of 116 stimulus-selective IT neurons during the "sample" period (100ms to 900ms after stimulus onset). Only a small number of IT neurons responded selectively during the delay period. For the PFC data, there were 67 stimulus-selective neurons during the sample period, and 32 stimulus-selective neurons during the immediately following "delay" period (300 to 1100 ms after stimulus offset, during which the monkey had to keep the category membership of the previously presented sample stimulus in mind, to compare it to a subsequently (at 1000 ms after stimulus offset) presented test stimulus [Freedman et al., 2001].

Figs. A.26 through A.28 show the BWI, CCI, and A_{ROC} distributions for the IT neurons (during the sample period — IT neurons tended to show much less delay activity than the PFC neurons), and the PFC neurons (during sample and delay periods, resp.).[2]

IT Comparing the view-tuned model unit data to the experimental IT data (Fig. A.29 and Fig. A.26), we observe a very good agreement of the BWI distributions of model units and IT neurons: Both are centered around zero and show a mean not significantly different from 0. Further, the ROC plots show very similar means, and — even more importantly — identical maxima (0.94). This shows that high ROC values can be obtained without any explicit category information, and moreover that the range of ROC values of experimental IT neurons are well compatible with those of view-tuned model units. There do appear to be some differences in the distribution of ROC values, with the experimental distribution having proportionally fewer neurons with intermediate ROC values.

Differences in the CCI distributions appear to be more substantial. However, the highest CCI in the model is greater than that of the experimental IT neurons, showing that model units can show similar degrees of category tuning as the experimental neurons.

Thus, in summary, the degree of category tuning of experimental IT neurons appears to be very well captured by the population of view-tuned model units. As model units were trained without any explicit category information, the agreement of experimental IT and model data suggest that the learning of IT neuron response properties can be understood as largely driven by shape similarities in input space, without any influence of explicit category information.

Comparison of model units *vs.* PFC neurons The PFC neurons show a BWI distribution with a positive mean significantly different from zero (sample period: 0.09, delay: 0.15), combined with higher average CCI values (sample: 0.21, delay: 0.21), with single neurons reaching values as high as 0.76 (sample and delay). Unlike in the IT case, this maximum value lies outside the range of CCI values of model units. Moreover, a positive average BWI of the magnitude found in the PFC data could only be obtained in the model with a significant number of border-tuned neurons. Such border-tuned units have very low CCI values. CCI values of PFC neurons are higher than those of IT neurons, however. Thus, the tuning properties of PFC neurons *cannot* be explained in the model by mere stimulus tuning alone, but seem to require the influence of explicit category information during training.

[2] For comparison with the model, the indices and ROC curves were calculated using a neuron's averaged firing rate (over at least 10 stimulus presentations) to each stimulus.

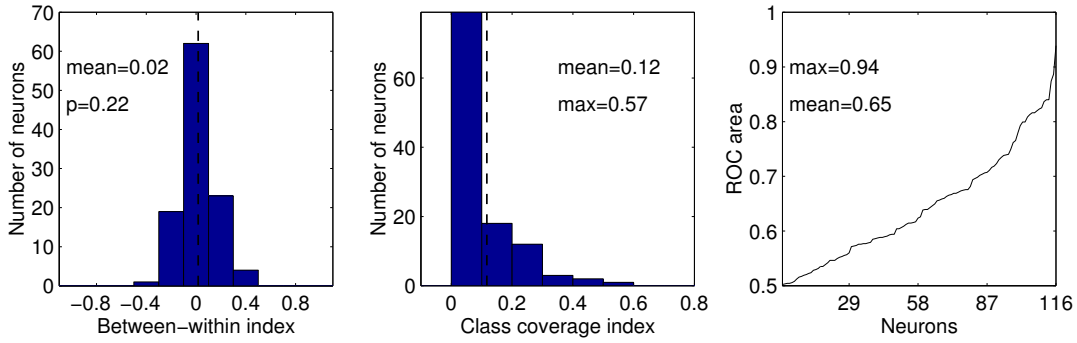


Figure A.26: Experimental IT data. The plots show the distribution of BWI (left), CCI (center) and ROC (right) area values.

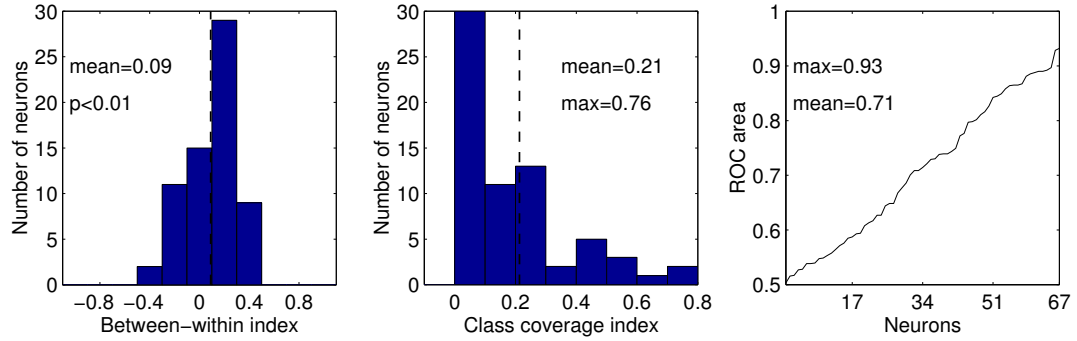


Figure A.27: Experimental PFC data (sample period). The plots show the distribution of BWI, CCI and ROC area.

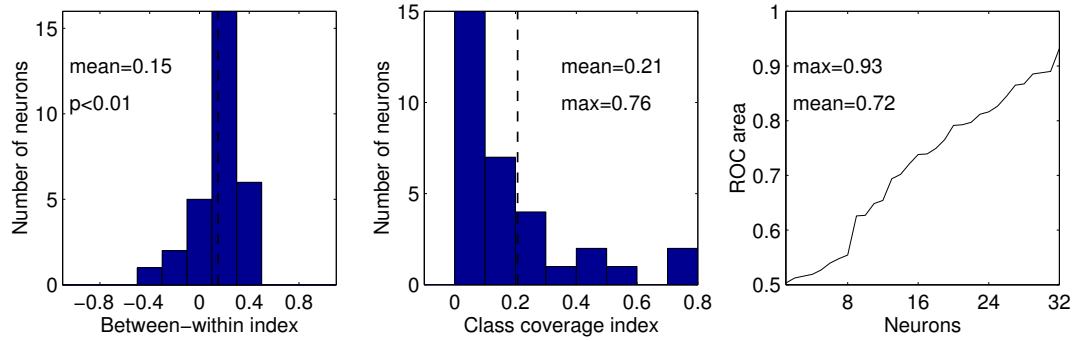


Figure A.28: Experimental PFC data (delay period). The plots show the distribution of BWI, CCI and ROC area.

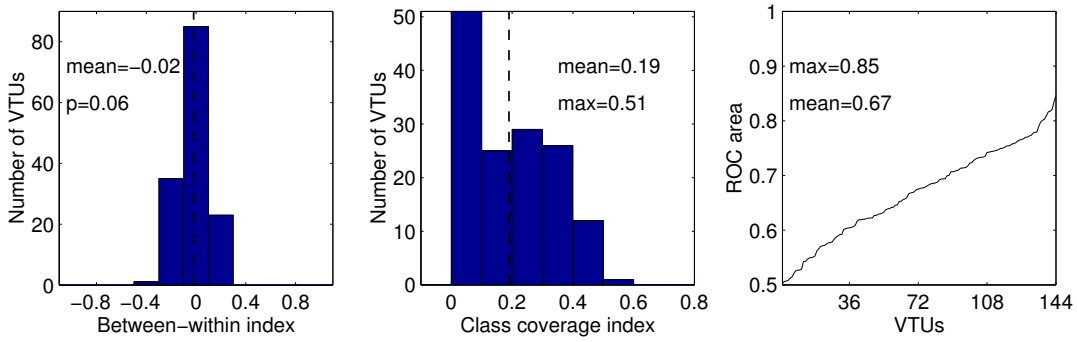


Figure A.29: Model IT data for $a = 32, \sigma = 0.2$. The plots show the distribution of BWI, CCI, and ROC area values.

Using the same analysis methods as in the experiment, we found that view-tuned model units showed tuning properties very similar to those of monkey IT neurons. In particular, as with IT cells in the experiment, we found that some view-tuned units showed "categorization-like" behavior, *i.e.*, very high ROC values. Most notably, this tuning emerged as a consequence of the shape-tuning of the view-tuned units, with no influence of category information during training. In contrast, the population of PFC neurons showed tuning properties that could not be explained by mere stimulus tuning. Rather, the simulations suggest the explicit influence of category information in the development of PFC neuron tuning.

These different response properties of neurons in the two brain areas, with IT neurons coding for stimulus shape and PFC neurons showing more task-related tuning, are compatible with a recent model of object recognition in cortex [Riesenhuber and Poggio, 1999b, 2000] in which a general object representation based on view- and object-tuned cells provides a basis for neurons tuned to specific object recognition tasks, such as categorization. This theory is also supported by data from another experiment in which different monkeys were trained on an identification and a categorization task, respectively, using the same stimuli [op de Beeck et al., 2001], and which found no differences in the stimulus representation by inferotemporal neurons of the monkeys trained on different tasks. On the other hand, a recent experiment [Sigala and Logothetis, 2002] reported IT neuron tuning emphasizing category-relevant features over non-relevant features (but no explicit representation of the class boundary, unlike in [Freedman et al., 2001]) in monkeys trained to perform a categorization task. Further studies comparing IT neuron tuning before and after training on a categorization task or even different tasks involving the same set of stimuli, and studies that investigate the possibility of top-down modulation from higher areas (*e.g.*, PFC) during task execution, will be needed to more fully understand the role of top-down task-specific information in shaping IT neuron tuning. The present study demonstrated the use of computational models to motivate and guide the analysis of experimental data. Clearly, the road ahead will equally require a very close interaction of experiments and computational work.

A.11 Biophysics details

This appendix provides more quantitative details about the models discussed in section 5. We start with a primer on synaptic transmission. Then, we describe the details of the nonspiking models, show some more detailed results for the maximum circuit and give a brief mathematical discussion of the circuit's input-output relation. Finally, we also provide details about the spiking circuit implementation.

A.11.1 Primer on the underlying biophysics of synaptic transmission

While a single neural signal is propagated between two neurons, it is transformed multiple times between two somata.

1. Starting with the membrane potential depolarization at the presynaptic cell's soma $\Delta V_{som,pre}$, this depolarization is propagated along the axon to the synaptic site causing a depolarization $\Delta V_{syn,pre}$. Usually this involves the creation of an action potential that travels down the axon to the presynaptic site and causes transmitter release. It seems possible, though, that an action potential is not necessary if the axon is short enough.
2. At the presynaptic site, the depolarization leads to Ca^{2+} influx and transmitter release T.
3. The transmitter binds to the receptors on the postsynaptic site and causes a conductance change g.
4. This conductance change leads to a current influx I and hence a synaptic membrane potential change $\Delta V_{syn,post}$.
5. The potential change travels through the dendrite, possibly invoking dendritic spikes, and finally causes a somatic membrane potential change $\Delta V_{som,post}$.

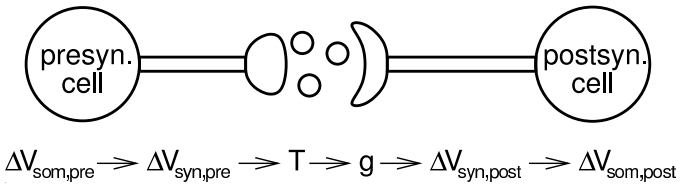


Figure A.30: Signal propagation between two somata.

The first goal is to find parametric relationships between the different stages:

$$\begin{aligned} \Delta V_{syn,pre} &= f_{ax}(\Delta V_{som,pre}) \\ T &= f_T(\Delta V_{syn,pre}) \\ g &= f_R(T) \\ \Delta V_{syn,post} &= f_V(g) \\ \Delta V_{syn,pre} &= f_{dend}(\Delta V_{syn,pre}) \end{aligned}$$

Notice that Katz and Miledi characterized the relation between $\Delta V_{syn,pre}$ and $\Delta V_{syn,post}$ at the neuromuscular junction as a monotonically increasing, nonnegative and bounded function [Katz and Miledi, 1970]. Reichardt *et al.* used a sigmoid to model this relation [Reichardt et al., 1983].

A.11.2 Non-spiking circuits

For the nonspiking circuits, we assume that the release of neurotransmitter at the synaptic terminal is graded and a continuous function of the graded depolarization of the cell.

We use a lumped model of a cell, i.e. each cell is modeled as a single compartment. Thus there are no delays or other cable effects for dendrites or axons at this point. Synapses are assumed to directly connect

the somata of the two involved cells. Each synapse is modeled in two stages: the transmitter release and the receptor binding. First the presynaptic potential is converted to a neurotransmitter concentration. We assume graded transmitter release, using a sigmoidal form of f_T (remember the amount of transmitter $T = f_T(V_{syn,pre})$). The relationship we used is

$$f_T(V_{pre}) = \frac{T_{max}}{1 + e^{-(V_{pre} - V_P)/K_P}}$$

where $T_{max} = 1mM$, $V_P = -10mV$ (half-activation point) and $K_P = 8mV$ (steepness). These values for V_P and K_P differ from the ones Destexhe *et al.* use [Destexhe et al., 1998] because their values were the result of fitting experimental data on synapses of spiking neurons. For spiking neuron synapses, a high threshold (corresponding to a high half-activation) and a steep transition seem reasonable. For graded potential synapses however, a lower threshold (and thus a lower half-activation point) and a less steep transition are desirable.

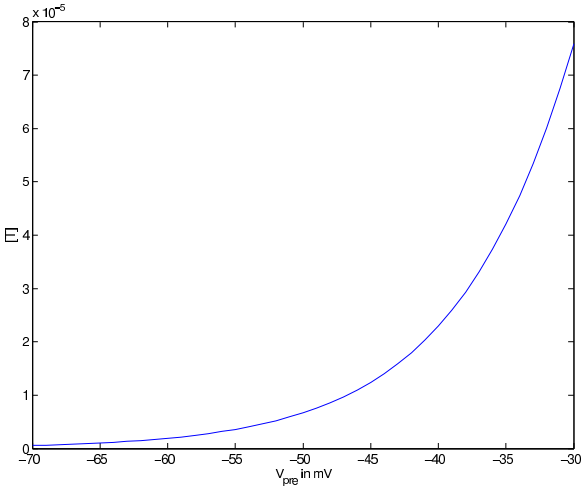


Figure A.31: Transmitter concentration T as a function of the presynaptic potential V_{pre}. The operating range is restricted to the left convex part of the sigmoid.

A simple way to model ionotropic receptors such as AMPA/Kainate and GABA$_A$ receptors is to assume simple two-state kinetics where the receptor is either open or closed based on the following diagram:

$$C + T \rightleftharpoons O,$$

described by the following first-order kinetic equation:

$$\frac{dr}{dt} = \alpha T(1 - r) - \beta r$$

where r is the proportion of open receptors and rate constants α and β taken from Destexhe *et al.* [Destexhe et al., 1998] to be $\alpha_{AMPA} = 1.1 \cdot 10^6 M^{-1} s^{-1}$, $\beta_{AMPA} = 190 s^{-1}$, $\alpha_{GABA_A} = 5 \cdot 10^6 M^{-1} s^{-1}$, $\beta_{GABA_A} = 6.6 s^{-1}$.

The input conductance can then be calculated as

$$g = r\bar{g}.$$

Finally, the membrane potential of the postsynaptic neuron is calculated based on all synaptic inputs as for integrate-and-fire neurons [Koch, 1998]:

$$C\frac{dV}{dt} = \sum g_{syn}(E_{syn} - V) + g_{leak}(E_{leak} - V).$$

The current results are based on AMPA as excitatory and GABA$_A$ as inhibitory receptors. The maximum conductance is varied for different simulations, but equal for all synapses of the same type. The reversal

potentials are $E_{leak} = -70mV$, $E_{AMPA} = 0mV$ and $E_{GABA_A} = -80mV$. The total leak conductance is set to be $g_{leak} = 6nS$, based on a cell with $30\mu m$ diameter and $0.2mS/cm^2$ leak conductance. The total cell capacitance is $30pF$, based on the same geometry and $1\mu F/cm^2$.

As input to the circuit, we set the membrane potential of the x units at different values. Based on that, the transmitter concentration and postsynaptic conductance for all synapses and the potential for the p, y and z cells are computed.

Max circuit results Figures A.32 through A.33 show the membrane potential of all units in the network (refer to Fig. 5.3 for a diagram of the network with the unit labels).

Figures A.34 and A.35 show the z unit traces for two individually presented stimuli (red traces) and the response to the combined stimulus (blue trace) as well as the sum and average of the two individual responses (solid and dotted green traces).

Figures A.36 through A.41 show the results (z unit traces) for different combinations of inputs, each figure for a different configuration, varying the network architecture (feedforward vs. feedback), maximal conductances and input onsets.

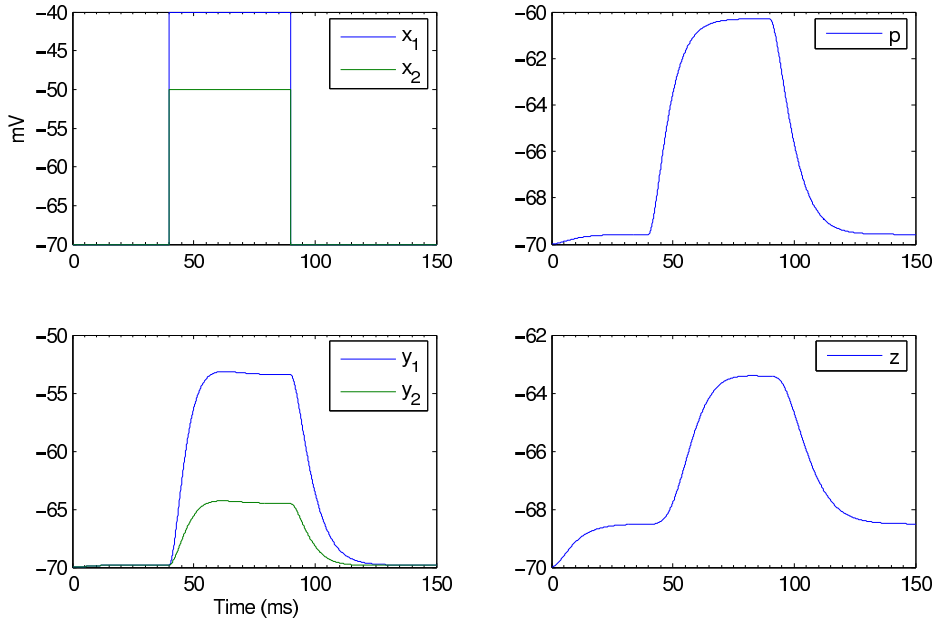


Figure A.32: Membrane potential for all cells in the network, feedforward model, $\bar{g}_{AMPA} = 20nS$, $\bar{g}_{GABA_A} = 20nS$. Stimulus length: 50ms, onset A:40ms, B:40ms

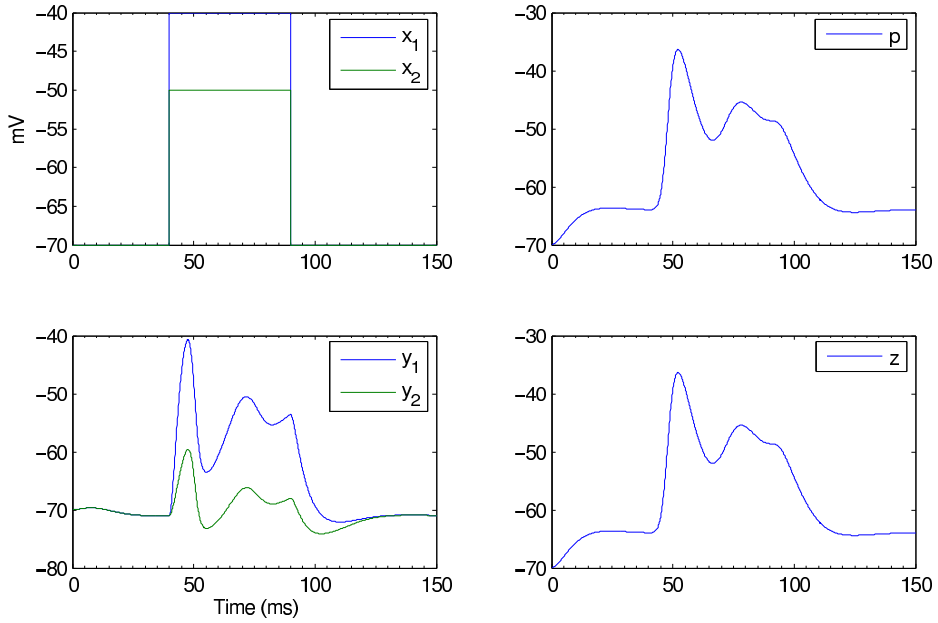


Figure A.33: Membrane potential for all cells in the network, feedback model, $\bar{g}_{AMPA} = 100nS$, $\bar{g}_{GABA_A} = 100nS$. Stimulus length: 50ms, onset A:40ms, B:40ms

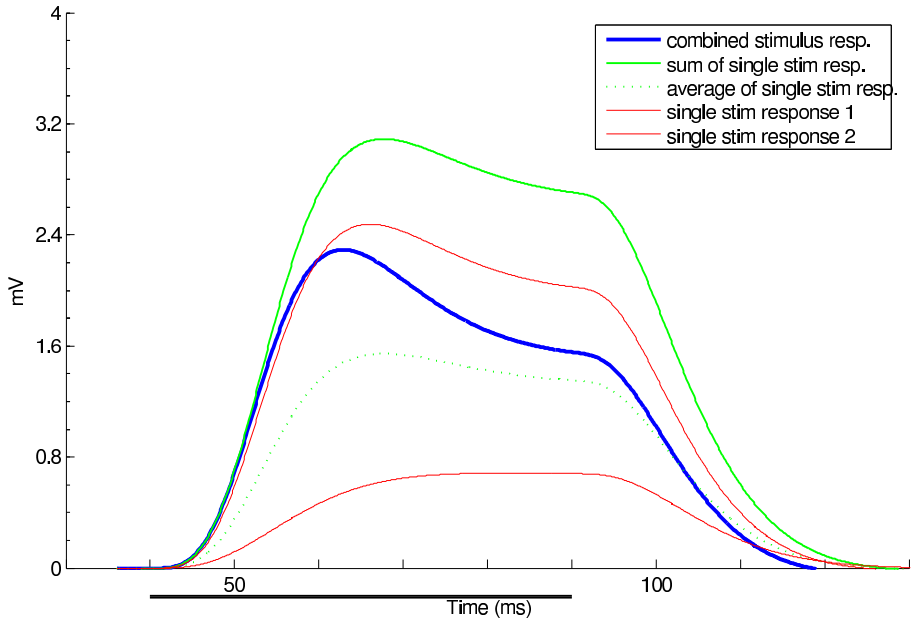


Figure A.34: Membrane potential for the z unit, feedforward model, $\bar{g}_{AMPA} = 20nS$, $\bar{g}_{GABA_A} = 20nS$. Stimulus length: 50ms, onset A:40ms, B:40ms

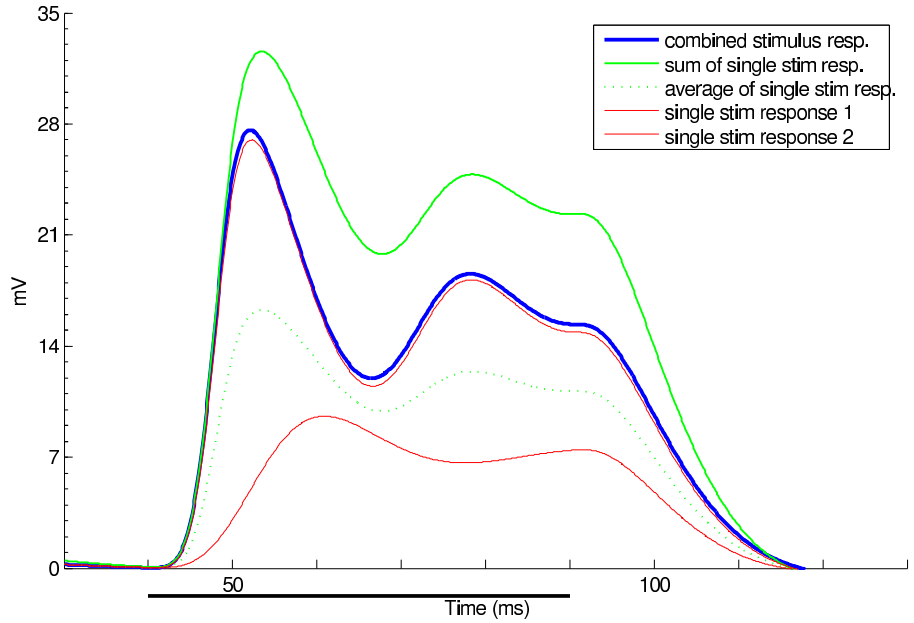


Figure A.35: Membrane potential for the z unit, feedback model, $\bar{g}_{AMPA} = 100nS$, $\bar{g}_{GABA_A} = 100nS$. Stimulus length: 50ms, onset A:40ms, B:40ms

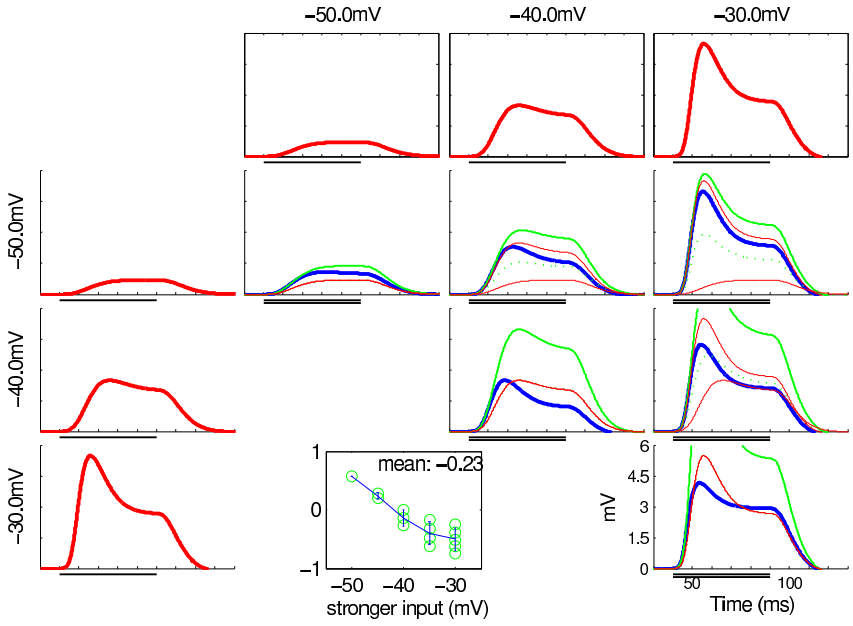


Figure A.36: Membrane potential for feedforward model, $\bar{g}_{AMPA} = 20nS$, $\bar{g}_{GABA_A} = 20nS$. Stimulus length: 50ms, onset A:40ms, B:40ms

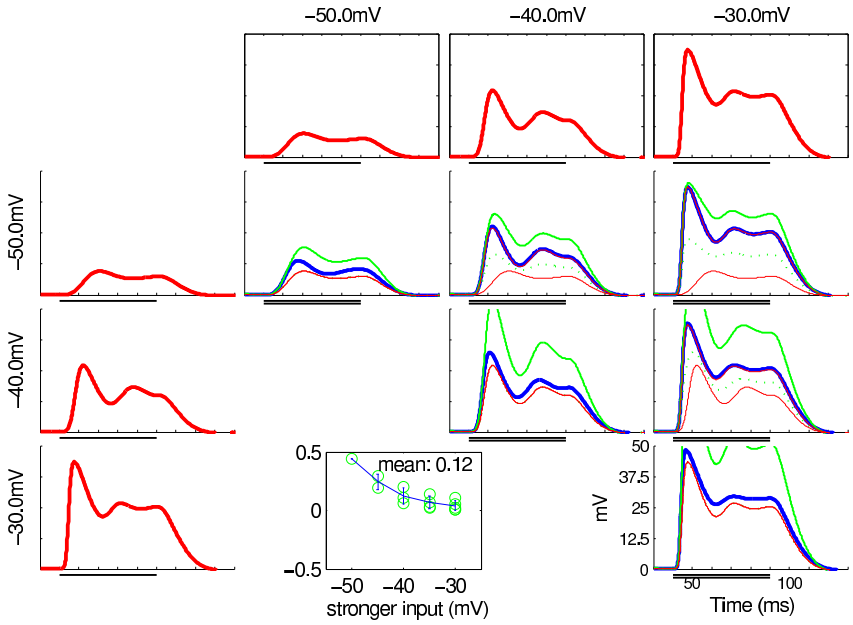


Figure A.37: Membrane potential for feedback model, $\bar{g}_{AMPA} = 20nS$, $\bar{g}_{GABA_A} = 20nS$. Stimulus length: 50ms, onset A:40ms, B:40ms

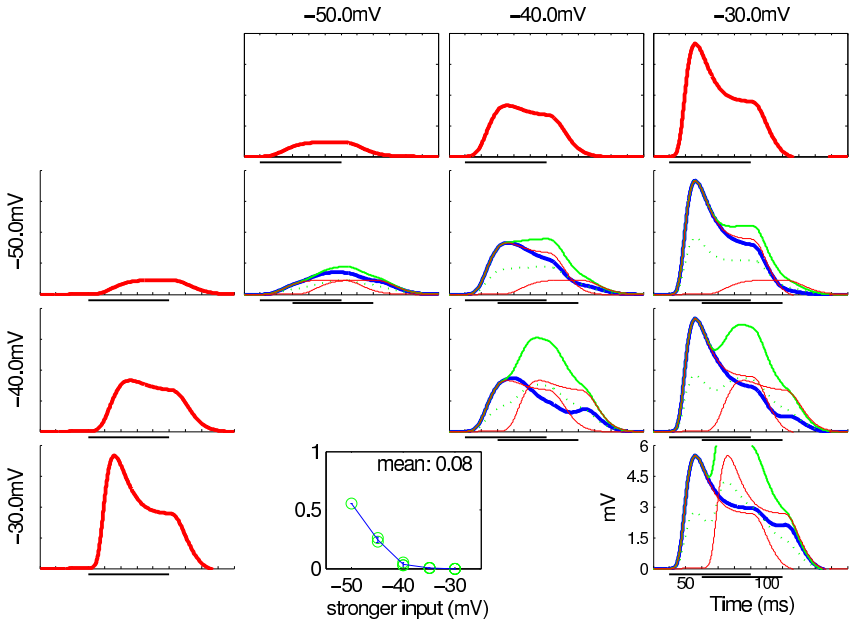


Figure A.38: Membrane potential for feedforward model, $\bar{g}_{AMPA} = 20nS$, $\bar{g}_{GABA_A} = 20nS$. Stimulus length: 50ms, onset A:40ms, B:60ms

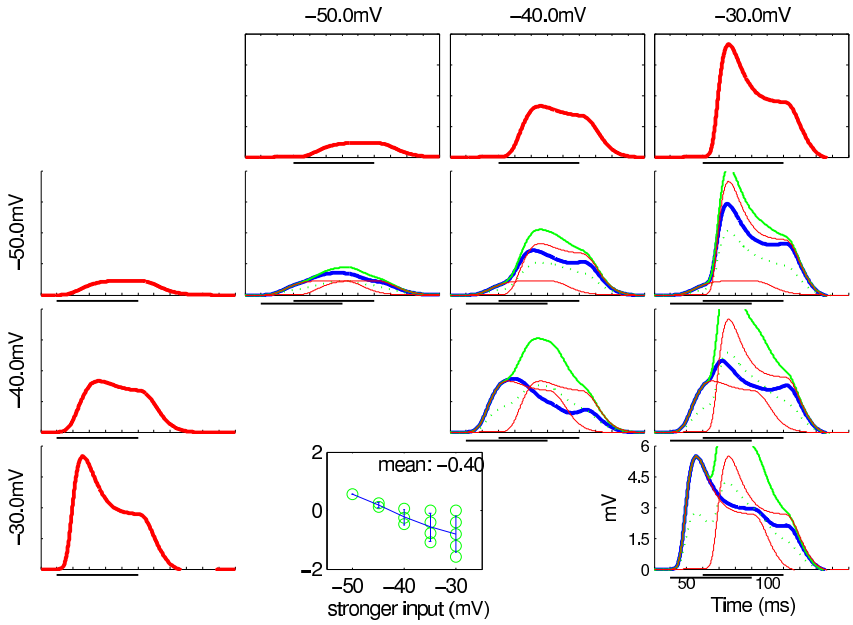


Figure A.39: Membrane potential for feedforward model, $\bar{g}_{AMPA} = 20nS$, $\bar{g}_{GABA_A} = 20nS$. Stimulus length: 50ms, onset A:60ms, B:40ms

Architecture variations Figures A.40 and A.41 show the response for the model with lateral inhibition for feedforward and feedback models. As before, the feedback circuit is more robust and closer to a max whereas the feedforward circuit's behavior depends more strongly on parameter settings and covers a range of sublinear regimes.

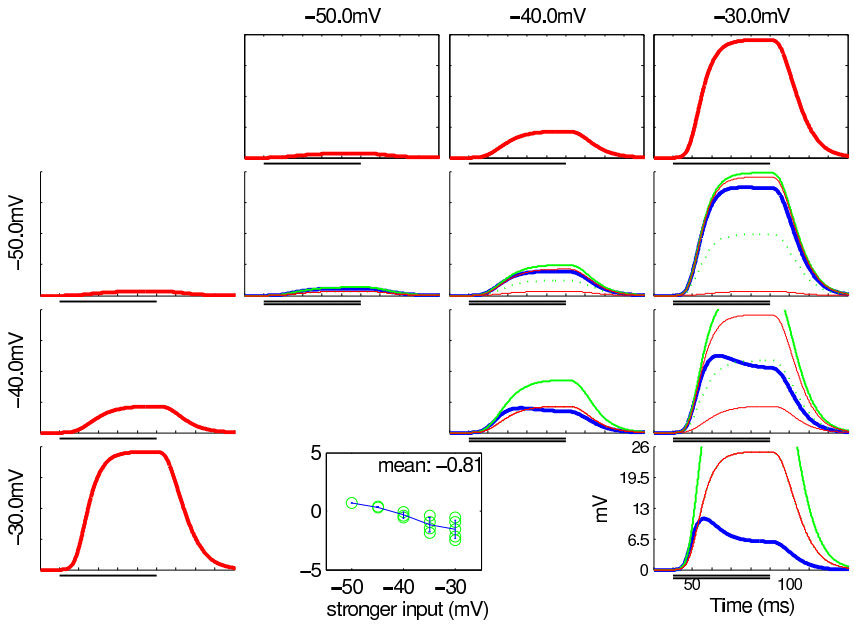


Figure A.40: Membrane potential for feedforward/lateral inhibition model, $\bar{g}_{AMPA} = 20nS$, $\bar{g}_{GABA_A} = 20nS$. Stimulus length: 50ms, onset A:40ms, B:40ms

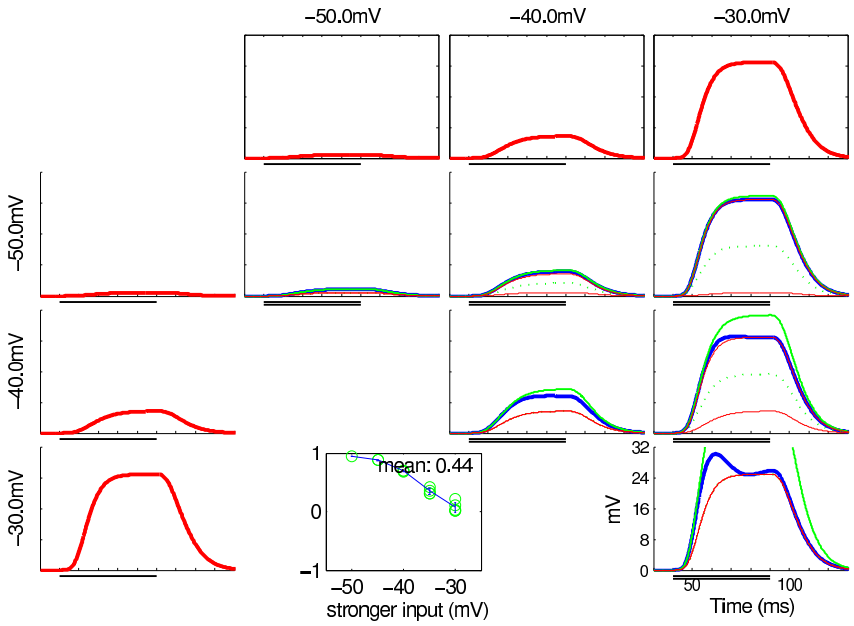


Figure A.41: Membrane potential for feedback/lateral inhibition model, $\bar{g}_{AMPA} = 20nS$, $\bar{g}_{GABA_A} = 20nS$. Stimulus length: 50ms, onset A:40ms, B:40ms

A.11.3 Spiking circuits

As shown in figure 5.7, each unit in the architecture is replaced by n equivalent cells which all receive the same input according to the circuit architecture. In the initial implementation, we chose each group to consist of $n = 50$ cells. Each cell is modeled as a single compartment leaky-integrate-and-fire cell. All excitatory cells are identical with membrane capacitance $C^e = 500pF$, resting potential $V_0^e = -60mV$, firing threshold $V_{th}^e = -45mV$, leak conductance $g_{leak}^e = 50nS$ and leak reversal potential $E_{leak}^e = -60mV$. Similarly, all inhibitory cells are identical with membrane capacitance $C^e = 200pF$, resting potential $V_0^e = -60mV$, firing threshold $V_{th}^e = -40mV$, leak conductance $g_{leak}^e = 50nS$ and leak reversal potential $E_{leak}^e = -60mV$. The difference in capacitance and firing threshold causes the fI-curve of the inhibitory cells to be steeper with a higher onset than for the excitatory cells (Fig.A.42.

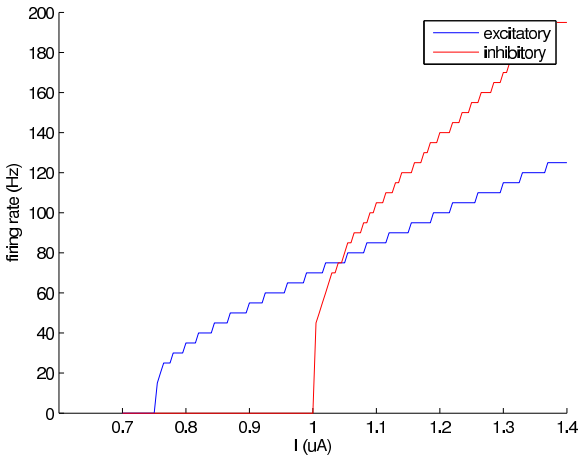


Figure A.42: f-I curves for excitatory and inhibitory leaky integrate-and-fire neurons in the model.

Each of the input x_i units, i.e. each of the equivalent cells, receives a square pulse current input of equal length (10ms) with normally distributed amplitude and onset.

When a cell fires, its membrane potential is reset to the resting potential and a 1ms square pulse of $1mM$ neurotransmitter is applied to the postsynaptic receptors, which are modeled identically to the nonspiking network.

For each arrow in figure 5.9 there are n equivalent cells and thus n synapses at each postsynaptic cell. Each of these synapses is modeled independently, contributing to the total synaptic conductance of the cell. In addition to the inputs from other cells in the circuit, each postsynaptic cell receives normally distributed background conductance input.

A.12 Brief discussion of some frequent questions

Here we briefly discuss some frequent questions that people raise about the model.

A.12.1 Connectivity in the model

The connectivity in the model is meant to follow some of the knowledge about anatomy of the connections in ventral visual cortex. However, important aspects of the connectivity in cortex are missing. The most notorious one is the degree of recurrent connections (see for example [Binzegger et al., 2004]). This is one of the important points where the theory needs to be extended (see 6.3) to incorporate feedback connections and study their functions in visual recognition.

A.12.2 Position invariance and localization

The invariance to scale or position achieved by the model is quite remarkable. However, there may be circumstances where you do not want invariance. One such case is the ability to localize an object within an image. Complete invariance would suggest that neurons cannot reveal the position of the object. While it is possible that object localization is performed in some other brain area (e.g. the dorsal visual stream), we observed that we can read out position information (at least coarsely) from the activity of a population of IT neurons [Hung et al., 2005a]. This is currently not incorporated into the model since the receptive field size of all S4 units is the same and is centered in the fovea. It does not seem too difficult, however, to change the input connectivity to S4 (or to the preceding layer) so that receptive fields vary in size and are centered at different positions across the field. These changes should easily allow to read out position information from the model.

A.12.3 Invariance and broken lines

Assuming that we have complete position invariance, there are many stimuli that might appear difficult to tell apart. One example includes differentiating between a continuous line, a broken line or other variants such as staggered lines (see examples in Figure A.43). However, it should be noted that there are multiple other units which may be able to differentiate these images (including, for example, some units which may prefer features oriented in a direction which is orthogonal to the line). We studied the responses of model units to a large sample of such broken line stimuli with 3 parameters: the bar size (b), the vertical separation (v) and the horizontal separation (h). These parameters are illustrated at the top of Figure A.43. By changing b, v and h, we considered 4 types of images: continuous lines, broken lines, staggered lines, staggered broken lines (Figure A.43. We then compared the responses of the model units for every possible pair of images (e.g. continuous vs. broken lines). The parameters used to generate the images covered a wide range of distances and separations but they did not include cases where the differences are smaller than the receptive field size of S1 units (e.g. hyperacuity was not considered here). In general, several (but not all) C2 units could easily differentiate among these different types of images. For example, for a particular set of parameters (bar size = 5 pixels, horizontal spacing = 20 pixels and vertical spacing = 15 pixels), most of the C2 units (1630 out of 2000 units) yielded the same response to the continuous line and the broken line. Most of the remaining 370 units yielded a stronger response to the continuous line (all images were constrained to show the same total number of white pixels). For the parameters explored here, at least some C2 units yielded differential responses even if the stimuli were equivalent for many or most of the C2 units. This does not exclude the possibility that units in earlier layers could detect collinearity or symmetry (see Section 6).

A.12.4 Configural information

There is no explicit account of configural information in the current version of the model. For example, in the case of recognizing a face, there is no unit which is specifically wired to code the distance between the eyes or the fact that the eyes should be above the nose. This would seem to suggest that it is possible to trick the model by randomly scrambling the different parts. However, random scrambling cuts across the receptive fields of the units and distorts the responses of units in the early layers, leading to a distortion

also in the responses at the level of the S4 units. This was illustrated for the case of paperclips in the work of Riesenhuber and Poggio [Riesenhuber and Poggio, 1999b]. How much S4 units are affected depends on the degree of scrambling; scrambling into smaller (random) boxes leads to a larger drop in the responses of the S4 units (compared to the responses to an unscrambled preferred image). Thus, even if configural information is not explicitly coded into the architecture and feature preferences, the model is not confused by scrambled images.

A.12.5 Invariance and multiple objects

It is a wrong intuition to assume that the response of a "complex" model unit to multiple objects always follows the response to the preferred stimulus in isolation, although it performs a maximum-like invariance operation. Response of the model unit is determined by the intricate interplay of the receptive field structure and the stimuli (in a similar manner as the above-mentioned situation of scrambling a stimulus image). If the co-occurring stimuli are sufficiently far away and do not fall within the receptive fields of the afferent units, the response will be invariant to the clutter. However, in many cases, the co-occurring stimuli will occupy the same receptive fields and produce non-trivial interference effects.

Some of the simulations and experiments described in this paper can actually be interpreted as such interference effects at multiple levels. Appendix A.7 shows that two co-occurring stimuli (two spots, in this case) modulate the response of a complex cell (C1 unit) depending on their relative configurations. In Section 4.2, V4 neurons (C2 units) show even more intricate pattern of interactions between two co-occurring stimuli (two spots or oriented bars). In Section 4.3.2, IT neurons (S4 or view-tuned units), although they are not strictly "complex" cells in the model, also display interference effects between multiple stimuli. In other words, the neural-level invariance operation does not necessarily carry onto the stimulus-level invariance, and for this reason, the effects of having multiple objects within the receptive field of a neuron or a model unit can be quite varied and unpredictable.

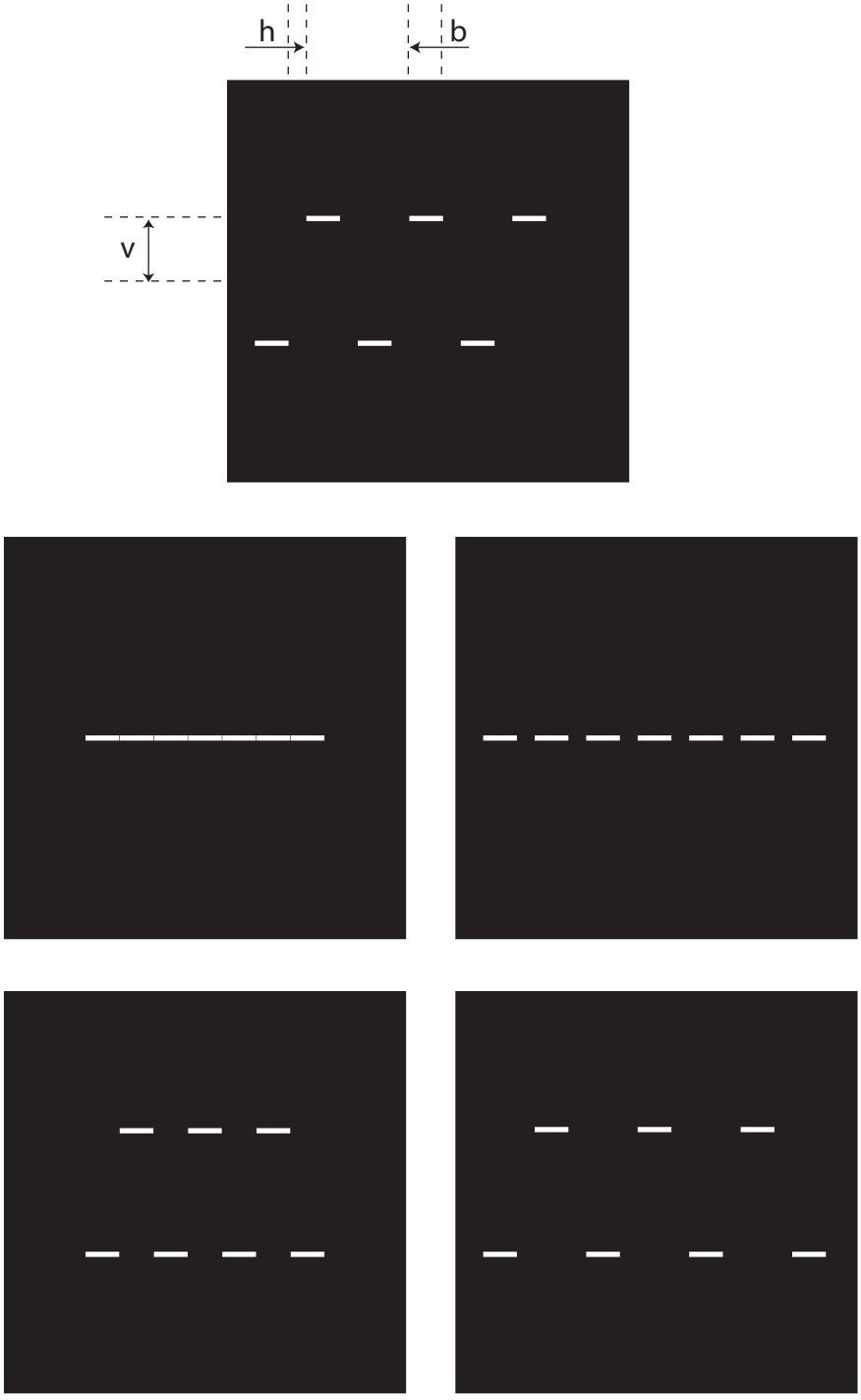


Figure A.43: Examples of images that might appear difficult to differentiate due to complete invariance of responses across the receptive field. At the top we show the main 3 parameters that we used to generate multiple such stimuli and study the responses of model units. These parameters include the bar size (b), the vertical separation (v) and the horizontal separation (h).

References

L.F. Abbott and W.G. Regehr. (2004). Synaptic computation. *Nature*, 431(7010):796–803. doi:10.1038/nature03010.

L.F. Abbott, J.A. Varela, K. Sen, and S.B. Nelson. (1997). Synaptic depression and cortical gain control. *Science*, 275:220–224.

N.C. Aggelopoulos, L. Franco, and E.T. Rolls. (2005). Object perception in natural scenes: encoding by inferior temporal cortex simultaneously recorded neurons. *Journal of Neurophysiology*, 93:1342–1357.

Y. Amit and M. Mascaro. (2003). An integrated network for invariant visual detection and recognition. *Vision Research*, 43(19):2073–2088.

C.H. Anderson and D.C. van Essen. (1987). Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proc. Nat. Acad. Sci. USA*, 84:6297–6301.

J.S. Bakin, K. Nakayama, and C.D. Gilbert. (2000). Visual responses in monkey areas V1 and V2 to three-dimensional surface configurations. *The Journal of Neuroscience*, 20(21):8188–8198.

H.B. Barlow. (1989). Unsupervised learning. *Neural Comp.*, 1:295–311.

S. Becker and G.E. Hinton. (1992). A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355:161–163.

H.O. op de Beeck, J. Wagemans, and R. Vogels. (2001). Inferotemporal neurons represent low-dimensional configurations of parametrized shapes. *Nat. Neurosci.*, 4:1244–1252.

R. Ben-Yishai, R.L. Bar-Or, and H. Sompolinsky. (1995). Theory of orientation tuning in visual cortex. *Proc. Nat. Acad. Sci. USA*, 92(9):3844–3848.

I. Biederman. (1987). Recognition-by-components: A theory of human image understanding. *Psych. Rev.*, 94:115–147.

S. Bileschi and L. Wolf. (2005). A unified system for object detection, texture recognition, and context analysis based on the standard model feature set. In *Proc. British Machine Vision Conference*.

T. Binzegger, R.J. Douglas, and K.A.C. Martin. (2004). A quantitative map of the circuit of cat primary visual cortex. *J. Neurosci.*, 24(39):8441–8453.

M.C. Booth and E.T. Rolls. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb. Cortex*, 8:510–523.

H. Bülthoff and S. Edelman. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Nat. Acad. Sci. USA*, 89:60–64.

C. Cadieu. (2005). Modeling shape representation in visual cortex area v4. Master's thesis, MIT, Cambridge, MA.

M. Carandini and D.J. Heeger. (1994). Summation and division by neurons in primate visual cortex. *Science*, 264:1333–1336.

M. Carandini, D.J. Heeger, and J.A. Movshon. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *J. Neurosci.*, 17:8621–8644.

J. Cavanaugh, W. Bair, and J. Movshon. (2002). Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *J. Neurophys.*, 88(5):2530–2546.

F.S. Chance, S.B. Nelson, and L.F. Abbott. (1998). Synaptic depression and the temporal response characteristics of V1 cells. *J. Neurosci.*, 18(12):4785–4799.

F.S. Chance, L.F. Abbott, and A.D. Reyes. (2002). Gain modulation from background synaptic input. *Neuron*, 35(4):773–782.

L. Chelazzi, J. Duncan, E.K. Miller, and R. Desimone. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *J. Neurophys.*, 80:2918–2940.

G.C. DeAngelis, J.G. Robson, I. Ohzawa, and R.D. Freeman. (1992). Organization of suppression in receptive fields of neurons in cat visual cortex. *J. Neurophysiol.*, 68(1):144–163.

G. Deco and E.T. Rolls. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, 44:621–644.

A. Delorme, G. Richard, and M. Fabre-Thorpe. (2000). Ultra-rapid categorisation of natural images does not rely on colour: A study in monkeys and humans. *Vision Research*, 40:2187–2200.

R. Desimone. (1991). Face-selective cells in the temporal cortex of monkeys. *J. Cogn. Neurosci.*, 3:1–8.

R. Desimone, T.D. Albright, C.G. Gross, and C. Bruce. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.*, 4(8):2051–2062.

R. Desimone, S.J. Schein, J. Moran, and L.G. Ungerleider. (1985). Contour, color and shape analysis beyond the striate cortex. *Vision Res.*, 25(3):441–452.

A. Destexhe, Z.F. Mainen, and T.J. Sejnowski. (1998). *Methods in Neuronal Modeling: From Ions to Networks*, Chapter 1: Kinetic Models of Synaptic Transmission, pages 1–26. MIT Press.

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

R.J. Douglas and K.A. Martin. (1991). A functional microcircuit for cat visual cortex. *J. Physiol. (Lond).*, 440: 735–69.

R.J. Douglas, C. Koch, M. Mahowald, K.A. Martin, and H.H. Suarez. (1995). Recurrent excitation in neocortical circuits. *Science*, 269(5226):981–5.

W. Einhäuser, C. Kayser, P. König., and K.P. Körding. (2002). Learning the invariance properties of complex cells from natural stimuli. *Eur. J. Neurosci.*, 15(3):475–486.

M.C.M. Elliffe, E.T. Rolls, and S.M. Stringer. (2002). Invariant recognition of feature combinations in the visual system. *Biol. Cyb.*, 86:59–71.

L. Fei-Fei, R. Fergus, and P. Perona. (2004). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR, Workshop on Generative-Model Based Vision*.

R. Fergus, P. Perona, and A. Zisserman. (2003). Object class recognition by unsupervised scale-invariant learning. In *CVPR*, volume 2, pages 264–271.

D. Ferster and K.D. Miller. (2000). Neural mechanisms of orientation selectivity in the visual cortex. *Annual Review of Neuroscience*, 23:441–471.

P. Földiák. (1991). Learning invariance from transformation sequences. *Neural Comp.*, 3:194–200.

P. Földiák. (1998). Learning constancies for object perception. In V. Walsh and J.J. Kulikowski, editors, *Perceptual Constancy: Why things look as they do*, pages 144–172. Cambridge Univ. Press, Cambridge, UK, 1998.

D.J. Freedman, M. Riesenhuber, T. Poggio, and E.K. Miller. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291:312–316.

D.J. Freedman, M. Riesenhuber, T. Poggio, and E.K. Miller. (2002). Visual categorization and the primate prefrontal cortex: Neurophysiology and behavior. *J. Neurophys.*, 88:930–942.

D.J. Freedman, M. Riesenhuber, T. Poggio, and E.K. Miller. (2003). Comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J. Neurosc.*, 415:5235–5246.

W.A. Freiwald, D.Y. Tsao, R.B.H. Tootell, and M.S. Livingstone. (2005). Complex and dynamic receptive field structure in macaque cortical area V4d. *Journal of Vision*, 4(8):184a.

K. Fukushima. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cyb.*, 36:193–202.

K. Fukushima. (1986). A neural network model for selective attention in visual pattern recognition. *Biol. Cyb.*, 55:5–15.

K. Fukushima, S. Miyake, and T. Ito. (1983). Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE Trans. on Systems, Man, and Cybernetics*, 13:826–834.

J.L. Gallant, C.E. Connor, S. Rakshit, J.W. Lewis, and D.C. van Essen. (1996). Neural responses to polar, hyperbolic, and cartesian gratings in area V4 of the macaque monkey. *J. Neurophys.*, 76:2718–2739.

T.J. Gawne. (2000). The simultaneous coding of orientation and contrast in the responses of V1 complex cells. *Exp. Brain Res.*, 133:293–302.

T.J. Gawne and J.M. Martin. (2002). Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. *J. Neurophys.*, 88:1128–1135.

M. Giese and T. Poggio. (2003). Neural mechanisms for the recognition of biological movements and action. *Nature Reviews Neuroscience*, 4:179–192.

C.G. Gross. (1998). *Brain Vision and Memory: Tales in the History of Neuroscience*. MIT Press.

C.G. Gross, C.E. Rocha-Miranda, and D.B. Bender. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *J. Neurophys.*, 35:96–111.

S. Grossberg. (1973). Contour enhancement, short term memory, and constancies in reverbarating neural networks. *Studies in Applied Mathematics*, 52:213–257.

R.H. Hahnloser, R. Sarpeshkar, M.A. Mahowald, R.J. Douglas, and H.S. Seung. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951. doi:10.1038/35016072.

J. Hawkins and S. Blakeslee. (2002). *On Intelligence*. Tomes Books, Holt, New York.

D.J. Heeger. (1993). Modeling simple-cell direction selectivity with normalized, half-squared, linear operators. *J. Neurophys.*, 70(5):1885–1898.

D.J. Heeger, E.P. Simoncelli, and J.A. Movshon. (1996). Computational models of cortical visual processing. *Proc. Nat. Acad. Sci. USA*, 93(2):623–627.

B. Heisele, T. Serre, M. Pontil, T. Vetter, and T. Poggio. (2001). Categorization by learning and combining object parts. In *Advances in Neural Information Processing Systems*, Vancouver.

J.K. Hietanen, D.I. Perrett, M.W. Oram, P.J. Benson, and W.H. Dittrich. (1992). The effects of lighting conditions on responses of cells selective for face views in the macaque temporal cortex. *Exp. Brain Res.*, 89: 157–171.

S. Hochstein and M. Ahissar. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36:791–804.

D. Hubel and T. Wiesel. (1965a). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophys.*, 28:229–89.

D.H. Hubel and T.N. Wiesel. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Phys.*, 160:106–154.

D.H. Hubel and T.N. Wiesel. (1965b). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophys.*, 28:229–289.

D.H. Hubel and T.N. Wiesel. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Phys.*, 195:215–243.

C. Hung, G. Kreiman, T. Poggio, and J. DiCarlo. (2005a). Fast read-out of object identity from macaque inferior temporal cortex. *Science*, in press.

C. Hung, G. Kreiman, T. Poggio, and J. DiCarlo. (2005b). Ultra-fast object recognition from few spikes. AI Memo 2005-022 / CBCL Memo 139, MIT AI Lab and CBCL, Cambridge, MA.

A. Hyvärinen and P.O. Hoyer. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vis. Res.*, 41(18):2413–2423.

E.R. Kandel, J.H. Schwartz, and T.M. Jessell. (2000). *Principles of Neural Science*. McGraw-Hill Companies, Inc.

B. Katz and R. Miledi. (1970). Further study of the role of calcium in synaptic transmission. *J. Phys.*, 207(3): 789–801.

C. Keysers, D.K. Xiao, P. Földiák, and D.I. Perrett. (2001). The speed of sight. *J. Cogn. Neurosci.*, 13:90–101.

U. Knoblich, D.J. Freedman, and M. Riesenhuber. (2002). Categorization in IT and PFC: model and experiments. AI Memo 2002-007 / CBCL Memo 216, MIT AI Lab and CBCL, Cambridge, MA.

E. Kobatake and K. Tanaka. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophys.*, 71:856–867.

E. Kobatake, G. Wang, and K. Tanaka. (1998). Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *J. Neurophys.*, 80:324–330.

C. Koch. (1998). *Biophysics of Computation: Information Processing in Single Neurons*. Oxford University Press, Oxford, England.

M. Kouh and T. Poggio. (2004). A general mechanism for cortical tuning: Normalization and synapses can create gaussian-like tuning. Technical Report AI Memo 2004-031 / CBCL Memo 245, MIT.

M. Kouh and M. Riesenhuber. (2003). Investigating shape representation in area V4 with HMAX: Orientation and grating selectivities. Technical Report AI Memo 2003-021 / CBCL Memo 231, Massachusetts Institute of Technology.

G. Kreiman. (2004). Neural coding: computational and biophysical perspectives. *Physics of Life Reviews*, 2: 71–102.

G. Kreiman, C. Hung, T. Poggio, and J. DiCarlo. (In press). Object selectivity of local field potentials and spikes in the inferior temporal cortex of macaque monkeys. *Neuron*.

I. Lampl, D. Ferster, T. Poggio, and M. Riesenhuber. (2004). Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. *J. Neurophys.*, 92:2704–2713.

Y. LeCun. (1988). A theoretical framework for back-propagation. In D. Touretzky, G. Hinton, and T. Sejnowski, editors, *Proc. of the 1988 Connectionist Models Summer School*, pages 21–28, CMU, Pittsburgh, PA. Morgan Kaufmann.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. (1998). Gradient-based learning applied to document recognition. *Proc. of the IEEE*, 86(11):2278–2324.

Y. LeCun, F.J. Huang, and L. Bottou. (2004). Learning methods for generic object recognition with invariance to pose and lighting. In *Proc. of CVPR'04*. IEEE Press.

P. Lennie. (1998). Single units and visual cortical organization. *Perception*, 27:889–935.

B. Leung. (2004). Component-based car detection in street scene images. Master's thesis, EECS, MIT.

O. Levy and I. Lampl. (2005). Amplification mechanism of cortical inhibitory circuits and its role in spatial suppression. In Prep.

L.F.Abbott, E.T.Rolls, and M.T.Tovee. (1996). Representational capacity of face coding in monkeys. *Cerebral Cortex*, 6:498–505.

F.F. Li, R. van Rullen, C. Koch, and P. Perona. (2002). Rapid natural scene categorization in the near absence of attention. *Proc. Nat. Acad. Sci. USA*, 99:9596–9601.

M. Livingstone and B. Conway. (2003). Substructure of direction-selective receptive fields in Macaque V1. *J. Neurophys.*, 89:2743–2759.

N.K. Logothetis and D.L. Sheinberg. (1996). Visual object recognition. *Ann. Rev. Neurosci.*, 19:577–621.

N.K. Logothetis, J. Pauls, H.H. Bülthoff, and T. Poggio. (1994). View-dependent object recognition by monkeys. *Curr. Biol.*, 4:401–413.

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

N.K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412:150–157.

J. Louie. (2003). A biological model of object recognition with feature learning. AI Memo 2003-009 / CBCL Memo 227, MIT.

D.G. Lowe. (1999). Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157.

N.A. Macmillan and C.D. Creelman. (1991). *Detection Theory: A User's Guide*. Cambridge University Press.

Y. Manor, F. Nadim, L.F. Abbott, and E. Marder. (1997). Temporal dynamics of graded synaptic transmission in the lobster stomatogastric ganglion. *J. Neurosci.*, 17(14):5610–5621.

J. Mariño, J. Schummers, D.C. Lyon, L. Schwabe, O. Beck, P. Wiesing, K. Obermayer, and M. Sur. (2005). Invariant computations in local cortical networks with balanced excitation and inhibition. *Nat. Neurosci.*, 8(2):194–201. doi:10.1038/nn1391.

D. Marr. (1982). *Vision : a computational investigation into the human representation and processing of visual information*. W.H. Freeman, San Francisco.

M. Maruyama, F. Girosi, and T. Poggio. (1991). Techniques for learning from examples: Numerical comparisons and approximation power. A.I. Memo 1290, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

M. Maruyama, F. Girosi, and T. Poggio. (1992). A connection between GRBF ane MLP. A.I. Memo 1291, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

B.W. Mel. (1997). SEEMORE: Combining color, shape and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Comp.*, 9(4):777–804.

E. Miller. (2000). The prefrontal cortex and cognitive control. *Nat. Rev. Neurosci.*, 1:59–65.

M. Missal, R. Vogels, and G.A. Orban. (1997). Responses of macaque inferior temporal neurons to overlapping shapes. *Cereb. Cortex*, 7:758–767.

U. Mitzdorf. (1985). Current source-density method and application in cat cerebral cortex: Investigation of evoked potentials and EEG phenomena. *Physiological Reviews*, 65:37–99.

Y. Miyashita. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, 335:817–820.

J. Movshon, I. Thompson, and D. Tolhurst. (1978). Receptive field organization of complex cells in the cat's striate cortex. *J. Phys.*, 283:79–99.

D. Mumford. (1996). On the computational architecture of the neocortex – II: The role of cortico-cortical loops. *Biol. Cyb.*, 66:241–251.

H. Nakamura, R. Gattass, R. Desimone, and L.G. Ungerleider. (1993). The modular organization of projections from areas V1 and V2 to areas V4 and TEO in macaques. *J. Neurosci.*, 13(9):3681–3691.

A. Oliva and A. Torralba. (In press). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*.

B.A. Olshausen and D.J. Field. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609.

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

A. Pasupathy and C.E. Connor. (1999). Responses to contour features in macaque area V4. *J. Neurophys.*, 82: 2490–2502.

A. Pasupathy and C.E. Connor. (2001). Shape representation in area V4: position-specific tuning for boundary conformation. *J. Neurophysiol.*, 86(5):2505–2519.

D.I. Perrett and M. Oram. (1993). Neurophysiology of shape processing. *Img. Vis. Comput.*, 11:317–333.

D.I. Perrett, P.A. Smith, D.D. Potter, A.J. Mistlin, A.S. Head, A.D. Milner, and M.A. Jeeves. (1984). Neurones responsive to faces in the temporal cortex: Studies of functional organisation, sensitivity to identity, and relation to perception. *Human Neurobiology*, 3:197–208.

D.I. Perrett, P.A. Smith, D.D. Potter, A.J. Mistlin, A.S. Head, A.D. Milner, and M.A. Jeeves. (1985). Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proc. of the Royal Society, London*.

D.I. Perrett, J.K. Hietanen, M.W. Oram, and P.J. Benson. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philos. Trans. Roy. Soc. B*, 335:23–30.

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

T. Poggio and S. Edelman. (1990). A network that learns to recognize 3D objects. *Nature*, 343:263–266.

T. Poggio, W. Reichardt, and W. Hausen. (1981). A neuronal circuitry for relative movement discrimination by the visual system of the fly. *Network*, 68:443–466.

M.C. Potter. (1975). Meaning in visual search. *Science*, 187:565–566.

R.P. Rao and D.H. Ballard. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects [see comments]. *Nat. Neurosci.*, 2(1):79–87.

W. Reichardt, T. Poggio, and K. Hausen. (1983). Figure-ground discrimination by relative movement in the visual system of the fly – II: Towards the neural circuitry. *Biol. Cyb.*, 46:1–30.

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

M. Riesenhuber and T. Poggio. (1999a). A note on object class representation and categorical perception. AI Memo 1679 / CBCL Memo 183, MIT AI Lab and CBCL, Cambridge, MA.

M. Riesenhuber and T. Poggio. (1999b). Hierarchical models of object recognition in cortex. *Nature Neurosc.*, 2:1019–1025.

M. Riesenhuber and T. Poggio. (2000). Models of object recognition. *Nat. Neurosci. Supp.*, 3:1199–1204.

M. Riesenhuber, M. Jarudi, S. Gilad, and P. Sinha. (2004). Face processing in humans is compatible with a simple shape-based model of vision. *Proc. Biol. Sci.*, 271:448–450.

D.L. Ringach. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque V1. *J. Neurophys.*, 88:455–463.

D.L. Ringach. (2004a). Haphazard wiring of simple receptive fields and orientation columns in visual cortex. *J. Neurophys.*, 92:468–476.

D.L. Ringach. (2004b). Mapping receptive fields in primary visual cortex. *J. Phys.*, 558.3:717–728.

E.T. Rolls and G. Deco. (2002). *Computational Neuroscience of Vision*. Oxford University Press, Oxford.

E.T. Rolls and M.J. Tovee. (1994). Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proc. Royal Society of London B*.

R. van Rullen and C. Koch. (2003a). Competition and selection during visual processing of natural scenes and objects. *Journal of Vision*, 3:75–85.

R. van Rullen and C. Koch. (2003b). Visual selective behavior can be triggered by a feed-forward process. *JCN*, 15:209–217.

R. van Rullen and S.J. Thorpe. (2001). Rate coding versus temporal order coding: What the retinal ganglion cells tell the visual cortex. *Neural Comp.*, 13(6):1255–1283.

V.R. de Sa and D.H. Ballard. (1998). Category learning through multi-modality sensing. *Neural Comp.*, 10 (5):1097–1117.

K. Sakai and Y. Miyashita. (2004). Neural organization for the long-term memory of paired associates. *Nature*, 354:152–155.

T. Sato. (1989). Interactions of visual stimuli in the receptive fields of inferior temporal neurons in awake macaques. *Exp. Brain Res.*, 74(2):263–271.

P.H. Schiller, B.L. Finlay, and S.F. Volman. (1976a). Quantitative studies of single-cell properties in monkey striate cortex I. Spatiotemporal organization of receptive fields. *J. Neurophysiol.*, 39(6):1288–1319.

P.H. Schiller, B.L. Finlay, and S.F. Volman. (1976b). Quantitative studies of single-cell properties in monkey striate cortex II. Orientation specificity and ocular dominance. *J. Neurophysiol.*, 39(6):1334–51.

P.H. Schiller, B.L. Finlay, and S.F. Volman. (1976c). Quantitative studies of single-cell properties in monkey striate cortex III. Spatial frequency. *J. Neurophysiol.*, 39(6):1334–1351.

O. Schwartz and E. Simoncelli. (2001). Natural signal statistics and sensory gain control. *Nat. Neurosci.*, 4 (8):819–825.

G. Sclar, J.H. Maunsell, and P. Lennie. (1990). Coding of image contrast in central visual pathways of the macaque monkey. *Vision Res.*, 30(1):1–10.

P. Series, J. Lorenceau, and Y. Fregnac. (2003). The silent surround of V1 receptive fields: theory and experiments. *Journal of Physiology - Paris*, 97:453–474.

T. Serre and M. Riesenhuber. (2004). Realistic modeling of simple and complex cell tuning in the HMAX model, and implications for invariant object recognition in cortex. AI Memo 2004-017 / CBCL Memo 239, MIT, Cambridge, MA.

T. Serre, M. Riesenhuber, J. Louie, and T. Poggio. (2002). On the role of object-specific features for real world object recognition. In S.W. Lee, H.H. Buelthoff, and T. Poggio, editors, *Proc. of Biologically Motivated Computer Vision*, Lecture Notes in Computer Science, New York. Springer.

T. Serre, L. Wolf, and T. Poggio. (2004). A new biologically motivated framework for robust object recognition. AI Memo 2004-026 / CBCL Memo 243, MIT, Cambridge, MA.

T. Serre, A. Oliva, and T. Poggio. (2005a). Ultra-rapid object categorization is compatible with a feedforward model of the visual pathway. *in prep.*

T. Serre, L. Wolf, S. Bileschi, and T. Poggio. (2005b). Robust object recognition with cortex-like mechanisms. In *Submitted*.

T. Serre, L. Wolf, and T. Poggio. (2005c). Object recognition with features inspired by visual cortex. In I.C.S. Press, editor, *Proc. of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, San Diego.

N. Sigala and N.K. Logothetis. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, 415:318–20.

R. Sigala, T. Serre, T. Poggio, and M. Giese. (2005). Learning features of intermediate complexity for the recognition of biological motion. In *ICANN 2005*.

D.J. Simons and R.A. Rensink. (2005). Change blindness: past, present and future. *Trends in Cognitive Science*, 9(1):16–20.

D.C. Somers, S.B. Nelson, and M. Sur. (1995). An emergent model of orientation selectivity in cat visual cortical simple cells. *J. Neurosci.*, 15(8):5448–5465.

M.W. Spratling. (2005). Learning view-point invariant perceptual representations from cluttered images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):753–761.

P.N. Steinmetz, A. Roy, P.J. Fitzgerald, S.S. Hsiao, K.D. Johnson, and E. Niebur. (2000). Attention modulates synchronized neuronal firing in primate somatosensory cortex. *Nature*, 404(6774):187–190.

J.V. Stone and A. Bray. (1995). A learning rule for extracting spatio-temporal invariances. *Network*, 6(3):1–8.

M.P. Stryker. (1991). Temporal associations. *Nature*, 354:108–109.

R.S. Sutton and A.G. Barto. (1981). Towards a modern theory of adaptive networks: expectation and prediction. *Psychol. Rev.*, 88:135–170.

R. Szulborski and L. Palmer. (1990). The two-dimensional spatial structure of nonlinear subunits in the receptive fields of complex cells. *Vis. Res.*, 30(2):249–254.

K. Tanaka. (1996). Inferotemporal cortex and object vision. *Ann. Rev. Neurosci.*, 19:109–139.

S. Thorpe. (2002). Ultra-rapid scene categorization with a wave of spikes. In *BMCV*, pages 1–15. Lecture Notes in Computer Science.

S.J. Thorpe and M. Fabre-Thorpe. (2001). Seeking categories in the brain. *Science*, 291:260–263.

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

A. Torralba, K.P. Murphy, and W.T. Freeman. (2004). Sharing features: efficient boosting procedures for multiclass object detection. In *CVPR*.

V. Torre and T. Poggio. (1978). A synaptic mechanism possibly underlying directional selectivity motion. *Proc. of the Royal Society London B*, 202:409–416.

S. Ullman, M. Vidal-Naquet, and E. Sali. (2002). Visual features of intermdediate complexity and their use in classification. *Nature Neuroscience*, 5(7):682–687.

L.G. Ungerleider and J.V. Haxby. (1994). 'What' and 'where' in the human brain. *Curr. Op. Neurobiol.*, 4: 157–165.

R.L.D. Valois, D.G. Albrecht, and L.G. Thorell. (1982a). Spatial frequency selectivity of cells in macaque visual cortex. *Vis. Res.*, 22:545–59.

R.L.D. Valois, E.W. Yund, and N. Hepler. (1982b). The orientation and direction selectivity of cells in macaque visual cortex. *Vis. Res.*, 22:531–44.

R. Vogels. (1999). Effect of image scrambling on inferior temporal cortical responses. *Neuroreport*, 10:1811–1816.

E. Wachsmuth, M.W. Oram, and D.I. Perrett. (1994). Recognition of objects and their component parts: Responses of single units in the temporal cortex of the macaque. *Cerebral Cortex*, 4:509–522.

G. Wallis and H.H. Bülthoff. (2001). Role of temporal association in establishing recognition memory. *Proc. Nat. Acad. Sci. USA*, 98(8):4800–4804.

G. Wallis and E.T. Rolls. (1997). A model of invariant object recognition in the visual system. *Prog. Neurobiol.*, 51:167–194.

G. Wallis, E.T. Rolls, and P. Földiák. (1993). Learning invariant responses to the natural transformations of objects. *International Joint Conference on Neural Networks*, 2:1087–1090.

D. Walther, T. Serre, T. Poggio, and C. Koch. (2005). Modeling feature sharing between object detection and top-down attention. *Journal of Vision*, 5(8):1041–1041. ISSN 1534-7362. URL `http://journalofvision.org/5/8/1041/`.

M. Weber, M. Welling, and P. Perona. (2000). Unsupervised learning of models for recognition. In *ECCV*, Dublin, Ireland.

H. Wersing and E. Koerner. (2003). Learning optimized features for hierarchical models of invariant recognition. *Neural Comp.*, 15(7):1559–1588.

L. Wiskott and T. Sejnowski. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Comp.*, 14(4):715–770.

A.J. Yu, M.A. Giese, and T. Poggio. (2002). Biophysiologically plausible implementations of the maximum operation. *Neural Comp.*, 14(12):2857–2881. doi:10.1162/089976602760805313.

H. Zhou, H.S. Friedman, and R. von der Heydt. (2000). Coding of border ownership in monkey visual cortex. *J. Neurosci.*, 20:6594–6611.

D. Zoccolan, D. Cox, and J.J. DiCarlo. (2005). Multiple object response normalization in monkey inferotemporal cortex. *J. Neurosci.*, 25(36):8150–8164.