Toward a More Biologically Plausible Model of Object Recognition

by

Minjoon Kouh

Submitted to the Department of Physics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2007

Author .....
Department of Physics
May 10, 2007

Certified by .....
Tomaso Poggio
Eugene McDermott Professor in the Brain Sciences and Human Behavior
Thesis Supervisor

Certified by .....
H. Sebastian Seung
Professor of Computational Neuroscience
Investigator of Howard Hughes Medical Institute
Thesis Supervisor

Accepted by .....
Thomas J. Greytak
Associate Department Head for Education

Toward a More Biologically Plausible Model of Object Recognition

by

Minjoon Kouh

Submitted to the Department of Physics
on May 10, 2007, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Rapidly and reliably recognizing an object (is that a cat or a tiger?) is obviously an important skill for survival. However, it is a difficult computational problem, because the same object may appear differently under various conditions, while different objects may share similar features. A robust recognition system must have a capacity to distinguish between similar-looking objects, while being invariant to the appearance-altering transformation of an object. The fundamental challenge for any recognition system lies within this simultaneous requirement for both specificity and invariance. An emerging picture from decades of neuroscience research is that the cortex overcomes this challenge by gradually building up specificity and invariance with a hierarchical architecture.

In this thesis, I present a computational model of object recognition with a feedforward and hierarchical architecture. The model quantitatively describes the anatomy, physiology, and the first few hundred milliseconds of visual information processing in the ventral pathway of the primate visual cortex. There are three major contributions. First, the two main operations in the model (Gaussian and maximum) have been cast into a more biologically plausible form, using monotonic nonlinearities and divisive normalization, and a possible canonical neural circuitry has been proposed. Second, shape tuning properties of visual area V4 have been explored using the corresponding layers in the model. It is demonstrated that the observed V4 selectivity for the shapes of intermediate complexity (gratings and contour features) can be explained by the combinations of orientation-selective inputs. Third, shape tuning properties in the higher visual area, inferior temporal (IT) cortex, have also been explored. It is demonstrated that the selectivity and invariance properties of IT neurons can be generated by the feedforward and hierarchical combinations of Gaussian-like and max-like operations, and their responses can support robust object recognition. Furthermore, experimentally-observed clutter effects and trade-off between selectivity and invariance in IT can also be observed and understood in this computational framework. These studies show that the model is in good agreements with a number of physiological data and provides insights, at multiple levels, for understanding object recognition process in the cortex.

Thesis Supervisor: Tomaso Poggio
Title: Eugene McDermott Professor in the Brain Sciences and Human Behavior

Thesis Supervisor: H. Sebastian Seung
Title: Professor of Computational Neuroscience
Investigator of Howard Hughes Medical Institute

Acknowledgments

My academic journey for the past 6 years has been sometimes exciting (for the joy of learning), other times frustrating, and overall humbling (for the privilege of being exposed to the decades of impressive scholarship by many great minds – of the past and the present – and learning about how some parts of the brain may work and how our perception and knowledge may be subject to inherent constraints and limitations).

I am truly grateful to my amazing mentor, Prof. Tomaso Poggio, who has given me much needed intellectual guidance throughout my PhD program and provided a wonderful research environment. He continues to show me how to "do" the science rigorously, while enjoying the process.

The members of the Center for Biological and Computational Learning have been wonderful companions. I thank Thomas Serre, Charles Cadieu, and Ulf Knoblich for their stimulating passions for science and research. Max Riesenhuber was my first mentor in the lab, who helped me to understand many important concepts in the beginning and taught me to pay attention to details. I am also grateful to Gabriel Kreiman for his good-natured, science-loving spirit and to Gadi Geiger for his wise presence (and for wonderful coffee). I also thank Adlar, Sasha, Sharat, Stan, Jake, Ethan, Tony, Huei-han, Cheston, and Jim, who have provided their impressive, keen intellect in the lab. I am very thankful to Mary Pat for all her administrative support and help.

I thank and pay my respect to many hard-working physiologists that I collaborated with. Davide Zoccolan especially has been a great colleague and friend. I appreciate the opportunities of working with Davide and Jim DiCarlo for the IT studies, and with Anitha Pasupathy and Ed Connor for the V4 studies. My gratitude also goes to Winrich Freiwald, Doris Tsao, and Marge Livingstone for their open and collaborative spirits. Winrich has been always supportive and educated me about the experimental techniques. I thank Nicole Rust for many interesting discussions about her work in MT.

More personally, I am thankful to my late-night friends, Alex and Helder (and Jim) in Building 46, for their support and for keeping me cheerful in numerous times. I thank Kyoungbong for his confidence in me and for the two memorable summers.

Occasional chances to play Pungmul (and many people I have met through it) have been an invaluable part of my graduate life. I am grateful for the precious friendship with Junghwan Park, Soo-Zin Kim, and Kyoung-Mun Shin, who have my highest respect and trust. I thank Kyoung-Mun especially for keeping me sane during my last year at MIT and for always being there and making me laugh.

I thank my lovely wife Yumi for her continuous support and trust in me, and for being the most wonderful companion. My wife and my son, Sukhwan, have been the joy of my life, and I am truly blessed by them. I am grateful to my mother-in-law for her continuous support. I am also thankful to my brother, a real physicist, who has inspired me to learn physics in college and actually taught me about the ocular motor muscles long before I became interested in the visual system (I don't know if he remembers this).

Most of all, I thank my parents who made everything possible. I am truly grateful for their love and support for us. This thesis and my other humble achievements would not have been possible without their sacrifices and brave choices in coming to a new country 17

years ago, and without their faith in us.

Contents

Chapter 1

Introduction

1.1 Goals and Approaches

Humans perform visual object recognition with phenomenal speed and robustness, far beyond the performance of the currently available computer vision systems. A 2-year old child easily outperforms even the most sophisticated, state-of-the-art computer vision systems in many essential visual tasks like object detection, recognition and categorization. Understanding the neural computations and representations of information in the cortex, underlying its remarkable efficiency, is one of the main goals of neuroscience.

Object recognition is a difficult computational problem because there are two conflicting requirements. On one hand, the recognition system needs to be selective and specific for different objects, which may in general look quite similar and share certain characteristic features (e.g., telling apart identical twins). On the other hand, the recognition needs to be invariant or tolerant to various appearance-altering transformations of an object (e.g., recognizing a person under disguise). Any learning process (e.g., generalizing an instance of a memory or rule to a new situation) may be considered as recognizing a consistent, invariant pattern within different contexts, or achieving an invariant selectivity for the relevant features.

The ventral pathway in the primate visual cortex, from the primary visual area V1 to inferior temporal cortex IT, is considered to mediate object recognition, based on its anatomy, physiology, and the lesion studies. Hence, this pathway is commonly called the "what pathway" [Ungerleider and Haxby, 1994]. Within the early visual areas along the pathway, such as V1, neurons tend to respond well to oriented bars or edges [Hubel and Wiesel, 1968]. Neurons in the intermediate visual areas, such as V2 and V4, are no longer tuned to oriented bars only, but to other forms and shapes of intermediate complexity [Desimone and Schein, 1987; Gallant et al., 1996; Pasupathy and Connor, 1999, 2001; Pollen et al., 2002; Freiwald et al., 2004]. Finally in the high visual areas like the inferior temporal cortex (IT), neurons are responsive to complex shapes like the image of a face or a hand [Kobatake and Tanaka, 1994; Gross et al., 1972; Logothetis et al., 1995; Hung et al., 2005; Zoccolan et al., 2005].

Two important trends are observed along the ventral pathway. The neurons are progressively tuned to more complex and specific shapes (*specificity* or *selectivity*), and their responses are increasingly tolerant to such transformations as translation and scaling (*invariance*). The balance of selectivity and invariance is essential for accomplishing visual tasks, because the brain sometimes has to distinguish between the similar-looking objects,

and other times recognize the same object under different viewing conditions.

It is an open question how the neural population in the ventral pathway represents the shape information and how selectivity and invariance properties are built up and traded off, in order to perform object recognition robustly. Many physiological studies have used different sets of visual stimuli in order to identify the underlying neural mechanisms. However, the nonlinear behaviors of the neurons in higher visual areas have made it difficult to determine the cortical computational mechanisms. Considering the infinite number of functions that can be fitted to the limited set of data points (given by the responses of a neuron to a set of test stimuli), studies that rely on *post hoc* function fitting are not feasible nor appropriate. Rather, it is essential to have an *a priori*, biologically plausible computational hypothesis of how complex neural computations are performed along the ventral pathway, a theory that provides testable predictions.

Therefore, a valuable investigative approach to a highly nonlinear system like cortex is computational modeling which provides a unifying framework to integrate experimental data and to make testable predictions. The computational model of Riesenhuber and Poggio [Riesenhuber and Poggio, 1999b] was originally proposed to account for the tuning properties (receptive field size, selectivity and invariance) of IT neurons, extending the classical suggestions by Hubel, Wiesel and others, and incorporating many standard findings and assumptions about the ventral pathway. Over the years, the model was extended and refined to incorporate more detailed comparisons with the neural data (see [Serre et al., 2005] and Fig. 3-1). The model is based on a hierarchical architecture of increasing invariance and selectivity, paralleling the neuronal shape tuning along the ventral pathway.

Using simulations of the model, this thesis attempts to develop a deeper understanding of the computational principles and mechanisms underlying the remarkable capabilities of the cortex, investigating how the complex neural networks in the primate visual cortex operate under their physical constraints and functional goals of performing object recognition. The use of a computational model allows for a quantitative measure of the system performance (i.e., how well it can recognize an object) and direct comparisons with the neurophysiological data. All the work in this thesis has been done in close collaborations with many experimentalists and theorists (which are not mutually exclusive categories), and the credits to them are acknowledged in the appropriate places throughout the thesis, as well as in the Acknowledgement section.

1.2 Preview

This thesis is composed of four parts.

Chapter 2 introduces a biologically plausible neural circuit, involving the mechanisms of divisive inhibition and polynomial-like monotonic nonlinearities. It is shown that such a circuit is capable of generating a variety of nonlinear neural responses, some of which can be characterized as Gaussian-like or maximum-like. Many different physiological data and modeling works have suggested that not only such a circuit is biologically feasible, but also it may constitute a canonical circuit, an elementary and universal unit of computation that may operate throughout the cortex.

Chapter 3 introduces a model of the ventral pathway [Riesenhuber and Poggio, 1999b; Serre et al., 2005], which is composed of hierarchical layers of these canonical neural circuits. This model belongs to a class of feedforward, hierarchical models of object recognition [Fukushima et al., 1983; Mel, 1997; Wallis and Rolls, 1997]. With detailed correspondences

with physiology and anatomy, the model describes the first few hundred milliseconds of visual processing along the feedforward path of the ventral stream in the primate visual cortex. The originally proposed principle operations, namely the Gaussian and the maximum operations [Riesenhuber and Poggio, 1999b], have been replaced by their approximations that are computable by the canonical circuit, and hence the model has been cast into a more biologically-plausible form.

In Chapter 4, the model units in the intermediate layer of the hierarchy have been used to model the tuning properties of the neurons in area V4, which also lies in the intermediate level of the ventral pathway. It is shown that a nonlinear combination of the oriented subunits (using a sequence of Gaussian-like and max-like operations) is capable of explaining the shape selectivity reported in different electrophysiological studies, such as [Pasupathy and Connor, 2001; Gallant et al., 1996; Desimone and Schein, 1987]. These neurons provide a rich (complete and redundant) dictionary of selective and invariant shape features, which can support even more complex and invariant shape tuning in the next layers of the hierarchical processing.

In Chapter 5, the model units in the top layer of the hierarchy have been used to model the tuning properties of the neurons in the inferior temporal (IT) cortex, which also lies in the top level of the ventral pathway, by comparing with three independent studies of the IT neurons [Logothetis et al., 1995; Hung et al., 2005; Zoccolan et al., 2007]. It is shown that the corresponding model units are selective to highly complex shape features and at the same time tolerant to position and size changes of the stimulus [Logothetis et al., 1995]. Therefore, they are capable of mediating invariant object recognition process, in agreements with the physiological data [Hung et al., 2005]. These model units are subject to the clutter effects, where introducing a cluttering object can modify the response to a single object. In a good agreement with the experimental data [Zoccolan et al., 2007], there is a wide range of clutter tolerance, as well as a wide range of shape selectivity within the population of the model units, and the selectivity and the tolerance (to clutter and to translation, size and contrast changes of the stimulus) properties are in general traded off within the population.

The work presented in this thesis corroborates the biological plausibility of a feedforward, hierarchical model of object recognition [Serre et al., 2005], based on its biologically plausible canonical neural circuitry and the close agreements with the physiological data at multiple levels of the ventral pathway.

Chapter 2

Canonical Neural Circuit for Cortical Nonlinear Operations

Two basic and complementary neural operations have been postulated over the last few years, suggested by experimental data across different cortical areas. A Gaussian-like operation yields a bell-shaped tuning over the patterns of activation of the presynaptic inputs. A max-like operation selects and transmits the response of the most active neural inputs. In this section, it will be shown that these two operations can be computed by the same "canonical" neural circuitry, involving divisive normalization and monotonic nonlinearities, for different values of a single parameter. The proposed canonical circuit combines the energy model [Adelson and Bergen, 1985] (for the weighted summation with polynomial nonlinearities) and the divisive normalization model, which are two widely (but usually separately) postulated neural mechanisms for explaining a variety of nonlinear response properties of cortical neurons. Some of this work will appear in [Kouh and Poggio, 2007].

2.1 Two Operations

Across the cortex, two broadly different types of neural responses have been observed. On one hand, many cortical cells produce strong activity for a certain "optimal" input pattern and decreasing activity for other inputs different from the optimal one [Hubel and Wiesel, 1962; Gallant et al., 1996; Pasupathy and Connor, 2001; Gross et al., 1972; Logothetis et al., 1995; Wilson et al., 2004; Rauschecker et al., 1995]. Such a behavior can be considered as a multidimensional tuning response, possibly generated by a Gaussian-like template-matching operation. In simple cases such as orientation tuning in the primary visual cortex, the observed neural tuning may arise from geometrical arrangement of the receptive fields of the afferent neurons [Hubel and Wiesel, 1962]. However, tuning behaviors along non-spatial dimensions may not be explained by the geometry of the input neurons alone. With hundreds and thousands of excitatory and inhibitory synapses, the input dimensionality of a cortical neuron is usually high, and the neural selectivity is expected to be determined by the "pattern" of its afferent inputs, not just by the sum of their activations. A formal description in terms of a Gaussian function is therefore quite natural.

On the other hand, neural responses in visual cortex are often tolerant to appearance-altering transformations of a stimulus, such as translation, rotation, and scaling [Hubel and Wiesel, 1962; Logothetis et al., 1995; Ito et al., 1995]. Some neurons are also tolerant to multiple cluttering stimuli within the receptive field: their responses may be described

with a maximum operation, which selects and transmits the strongest response among a pool of several scalar inputs [Lampl et al., 2004; Gawne and Martin, 2002]. For simple image transformations like translation on a uniform background, a summation operation is sufficient to explain the tolerance properties, but for more difficult cases like clutter, the summation alone is not, as each object in the scene would contribute to the sum and provide a highly perturbed and intolerant output. As a result, the summation operation can "hallucinate" or confuse the presence of a target object with the presence of several less optimal objects. In contrast, a max-like operation can provide clutter tolerance more effectively, as well as the tolerance to other object transformations [Riesenhuber and Poggio, 1999b].

Although their specific mathematical forms should not be taken too literally, the following idealized functions implement those neural operations, which may underlie the described neural properties:

$$\text{Gaussian:} \quad y(\vec{x}) = \quad e^{-|\vec{x}-\vec{w}|^2/2\sigma^2}, \tag{2.1}$$

$$\text{Maximum:} \quad y(\vec{x}) = \quad \max_i(x_i), \tag{2.2}$$

where a scalar value y corresponds to the strength of the neural response or a monotonic function of it (e.g., the number of spikes or firing rate) for a given input $\vec{x}$, which corresponds to the activity of the presynaptic afferent neurons, and where x_i is the i-th component of the vector $\vec{x}$. In the Gaussian function, $\vec{w}$ denotes the optimal input pattern that produces the highest output, and σ determines the sharpness or sparseness of the tuning.

Importantly, these equations refer to the neural operations that act on the responses of the afferent neurons, not directly on the external stimuli. Because it is difficult to measure and manipulate the input and output of a neural circuit, the observed Gaussian-like and max-like response profiles to the experimental stimuli provide only indirect evidence for such neural-level operations, which are, hence, a hypothesis that needs to be tested with detailed biophysical experiments.

2.1.1 Gaussian-like Tuning

Especially in the sensory areas of the cortex, many neurons respond strongly to some stimuli, but weakly to others, as if they are tuned to certain patterns of activity of their inputs. For example, some of the neurons in the primary visual cortex show Gaussian-like tuning in multiple dimensions, such as orientation, spatial frequency, direction and velocity [Hubel and Wiesel, 1962]. Further along the ventral pathway of the primate visual cortex, some neurons in area V4 show tuned responses to different types of gratings or contour features [Gallant et al., 1996; Pasupathy and Connor, 2001], and inferior temporal neurons are tuned in a bell-shaped way to more complex shapes such as the view of an object [Gross et al., 1972; Logothetis et al., 1995]. Some neurons in MT are selective for the motion of a global pattern, displaying a Gaussian-like tuning behavior to a set of specific motion components of a moving stimulus [Rust et al., 2006]. In other sensory modalities, Gaussian-like neural selectivities are also reported (e.g., for the olfactory neurons in flies [Wilson et al., 2004] and auditory neurons in primates [Rauschecker et al., 1995]). In the motor system, the activity of a spinal cord neuron can be regarded as being tuned to a particular pattern of force fields or limb movements [Poggio and Bizzi, 2004]. Some neurons in hippocampus, known as place cells, are found to be selective for the spatial position of an animal [O'Keefe, 1976]. In higher cortical areas, such as the prefrontal cortex or lateral intraparietal cortex

[Freedman et al., 2003; Freedman and Assad, 2006], some neurons are tuned to the learned categories, suggesting (arguably) that the neural selectivity is also deeply involved in such cognitive functions of learning and categorization. The tuned response of a neuron can be sharp, broad, sparse, or distributed [Kreiman, 2004], but overall, the exquisite neural selectivity is believed to be one of the major computational strategies for representing and encoding information in the cortex.

From a theoretical side, it has been demonstrated that radial basis function (RBF) networks using a Gaussian kernel can learn effectively from "small" training sets and generalize the input-output mapping to a new set of data [Poggio, 1990; Poggio and Bizzi, 2004]. Such an architecture and its generalization properties may be the basis of the stimulus-specific tuning behaviors observed in the cortical networks. Notice that an advantage of the networks with the Gaussian-like tuning units (as opposed to a perceptron-like network with the sigmoid neural units only) is the speed and ease of learning the parameters in the network [Moody and Darken, 1989; Poggio and Girosi, 1989]. Learning is even easier for a normalized RBF network, because the synaptic weights from the Gaussian units to the output are simply the values of the function at the example points [Girosi et al., 1995].

2.1.2 Max-like Operation

In the primary visual cortex, many neurons (probably a subset of the "complex" cells) show tuned, selective responses to different orientations of a bar or Cartesian grating stimulus, while being tolerant to small perturbations in the stimulus location [Hubel and Wiesel, 1962]. Such translation invariance properties are found in other parts of the primate visual cortex [Gallant et al., 1996; Pasupathy and Connor, 2001; Logothetis et al., 1995; Ito et al., 1995]. As proposed by Fukushima and others [Hubel and Wiesel, 1962; Fukushima et al., 1983; Perrett and Oram, 1993; Riesenhuber and Poggio, 1999b], a plausible feedforward mechanism for invariance is to pool from the afferent cells tuned to the transformed versions of a stimulus (e.g., translation, scale, and rotation). For example, a complex cell in the primary visual cortex may pool from the simple cells with the same orientation selectivity but with receptive fields at the neighboring positions, so that its output would still be orientation selective but invariant to the exact spatial locations of the stimuli [Hubel and Wiesel, 1962]. In [Riesenhuber and Poggio, 1999b], a maximum operation, or its soft-max approximation, was suggested for the pooling. Similarly, by pooling together the afferent neurons tuned to the different views of an object, it may be possible to generate a view-invariant output [Perrett and Oram, 1993] (though in this case, [Poggio and Edelman, 1990] suggested a Gaussian-based RBF architecture).

A particularly difficult, yet important form of invariance is the tolerance to clutter, since normally objects do not appear in isolation, but in a cluttered context or scene (e.g., listening to a particular speaker in a cocktail party or looking for a key on a cluttered desk). A few recent physiological experiments have reported that when multiple stimuli are simultaneously presented in such a way that they are likely to stimulate different inputs to the recorded neurons, these neurons produce a max-like response or selection: for example, complex cells in the primary visual cortex [Lampl et al., 2004] and some neurons in area V4 [Gawne and Martin, 2002] or MT [Nowlan and Sejnowski, 1995]. (An average-operation, which produces a sublinear output like the maximum, may work somewhat similarly.)

In general it is difficult to control that, for example, two bar stimuli in the retina will activate different inputs to a complex cell in V1 or a cell in V4, because of the several intervening stages of converging and diverging neural connectivities in the cortex. Thus it

is not surprising that a number of experiments with multiple stimuli do not show a max-like behavior, but a variety of different response patterns: the complex cells in V1 [Movshon et al., 1978; Livingstone and Conway, 2003], V2 and V4 neurons [Reynolds et al., 1999; Freiwald et al., 2004]. Especially when the neurons themselves may not be performing a max-like operation (possibly, the neurons in the inferior temporal cortex [Zoccolan et al., 2005]), the max-like clutter tolerant behaviors will not readily observable, either.[1]

2.2 Neural Circuits

While there has been a number of experimental and theoretical motivations for the plausibility of the Gaussian-like and max-like neural operations, it remains a puzzle how the cortex may actually implement them. In this section, biologically plausible neural circuits, capable of computing close approximations to the idealized operations (Eq. 2.1 and 2.2) are proposed. The most general form of computations that these circuits can perform is summarized in Table 2.1, along with the associated functions that can approximate the Gaussian and the maximum, respectively. Only a static approximation of the neural circuit is considered. A more detailed description would involve the dynamics of realistic synapses and spiking neurons.

2.2.1 Circuit A and B: Divisive Normalization

Both circuits A and B in Fig. 2-1 perform a weighted summation over normalized (or gain-controlled) inputs. The key element is the divisive normalization of the input, which may be accomplished by feedforward or lateral shunting inhibitions, as exemplified by the two circuits in the figure respectively (see [Serre et al., 2005; Reichardt et al., 1983; Carandini

[1]The change of neural response in the presence of multiple stimuli within the receptive field of a neuron is found not just in vision, but also in other sensory modalities, although these effects may be called by different names, such as masking, interference, crowding, competition, etc. The clutter effects include suppression, facilitation, average, and other modulatory behaviors, which may be explained by the following arguments: First, the proposed max-like operation occurs at the level of the neural responses, not at the level of the stimuli–that is, the response to the multiple stimuli by a "max neuron" will not necessarily be the maximum response to the individual stimulus, especially if the max operation occurs after multiple processing stages in the cortical hierarchy. For example, suppose two bars are used to probe the receptive field of an orientation-selective complex cell, which is assumed to be pooling from several simple cells with the same orientation selectivity, in the primary visual cortex. If these two stimuli separately activate different afferent neurons, the maximum operation can produce a clutter-invariant output. Alternatively, they may also appear together within the receptive field of an afferent neuron, producing an interference and perturbing the afferent responses (i.e., $x_{A+B} \neq x_A$ nor x_B, where x_S denotes the response of an input neuron to some stimulus S). Additionally, it is also possible that the less optimal stimulus, which would produce a weak response by itself, may actually activate some afferent neurons more strongly (i.e., $x_A < x_B$, even though $y_A > y_B$, where y_S is the response of the efferent neuron to some stimulus S). In these likely cases, even if the underlying operation was in fact the maximum, the output would not be the maximum response to the individual stimulus (i.e., $y_{A+B} \neq \max(y_A, y_B)$, even if $y = \max(\vec{x})$). Hence, varying degrees of clutter tolerance are expected, for example, from the complex cells in the primary visual cortex [Movshon et al., 1978; Livingstone and Conway, 2003; Lampl et al., 2004]. The maximum operation is more likely to be observable in cases where multiple stimuli tend not to co-occupy the receptive fields of the afferent neurons (e.g., as pointed out in [Gawne and Martin, 2002], more max-like interactions are observed when two stimuli are presented farther apart, as compared to the more suppressive interactions reported in [Reynolds et al., 1999]). Second, the neurons performing a Gaussian-like tuning operation can also show a variety of response modulations, depending on exactly how multiple stimuli change the afferent activities. The perturbation by the clutter stimuli may produce an input pattern that is closer to (or farther away from) the optimal, in which case the response will be facilitated (or suppressed).

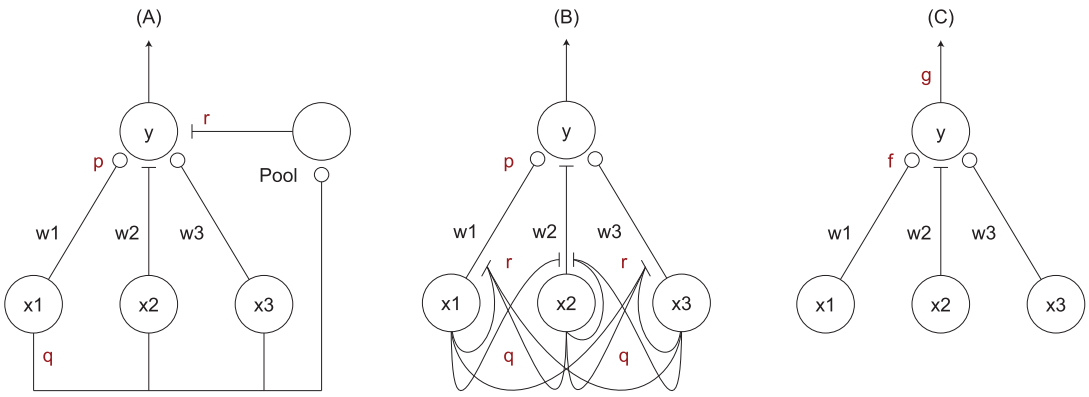


Figure 2-1: Biologically plausible neural circuits for implementing both Gaussian-like and max-like operations. The excitatory and inhibitory connections are denoted by an open circle and a bar, respectively. Circuit (A) performs divisive normalization with feedforward shunting inhibition. The same operation may be computed by lateral inhibition as in circuit (B). Such circuits have been proposed and studied earlier in many different contexts [Grossberg, 1973; Reichardt et al., 1983; Carandini and Heeger, 1994; Nowlan and Sejnowski, 1995; Lee et al., 1999; Yuille and Grzywacz, 1989]. Instead of shunting inhibition, the normalization, or gain control, mechanism may also rely on the inherent biophysics of dendritic integration [Borst et al., 1995]. In the limit of very weak divisive normalization, the circuits A and B reduce to a much simpler circuit C, which performs a weighted summation. Some of the synaptic weights in these circuits may be negative (e.g., w_2), possibly mediated by interneurons (not drawn). Monotonic nonlinearities in the circuits are denoted by p, q, and r. See Appendix B.1.

Operation	Circuit A or B	Circuit C
Canonical	$y = \dfrac{\sum_{j=1}^{n} w_j \; {x_j}^p}{k + \left(\sum_{j=1}^{n} {x_j}^q \right)^r}$ (2.3)	$y = g\left(\sum_{j=1}^{n} w_j \; f(x_j) \right)$ (2.4)
Gaussian-like	$y = \dfrac{\sum_{j=1}^{n} w_j \; x_j}{k + \sqrt{\sum_{j=1}^{n} {x_j}^2}}$ (2.5)	$y = \tanh\left(\sum_{j=1}^{n} w_j \; x_j \right)$ (2.6)
Max-like	$y = \dfrac{\sum_{j=1}^{n} {x_j}^{q+1}}{k + \left(\sum_{j=1}^{n} {x_j}^q \right)}$ (2.7)	$y = \log\left(\sum_{j=1}^{n} e^{x_j} \right)$ (2.8)

Table 2.1: Potential computations that can be performed by the neural circuits in Fig. 2-1 at their steady states. The canonical form is given in the first row, and the corresponding examples of the Gaussian-like and max-like operations are given in the next two rows. The responses of the input neurons are denoted by $\vec{x}$, and their synaptic weights, by $\vec{w}$. The summation in the denominator may also include different synaptic weights as in [Fukushima et al., 1983; Schwartz and Simoncelli, 2001; Rust et al., 2006]. Possible monotonic nonlinearities due to different synaptic efficacies are approximated by p, q, r, f and g.

and Heeger, 1994; Heeger, 1993; Grossberg, 1973; Fukushima et al., 1983] and Appendix B.1, but also [Holt and Koch, 1997]).

Such circuits have been proposed and studied in the past in various contexts: Reichardt et al. considered the forward and recurrent shunting inhibition circuits for gain control and detection of motion discontinuities by the fly visual system [Reichardt et al., 1983]. A similar normalization mechanism was used to explain contrast-dependent, saturating neural responses [Carandini et al., 1997; Heeger, 1993; Carandini and Heeger, 1994] and center-surround effect [Cavanaugh et al., 2002; Series et al., 2003] in the primary visual cortex. A divisive normalization scheme can increase the independence of the neural responses, despite the dependencies in the inputs [Schwartz and Simoncelli, 2001] and help to stabilize recurrent neural networks [Chance and Abbott, 2000]. The divisive normalization is also likely to constitute an important part of motion selectivity in MT [Nowlan and Sejnowski, 1995; Rust et al., 2006] and of attentional mechanism [Reynolds et al., 1999; Lee et al., 1999]. In [Yu et al., 2002; Riesenhuber and Poggio, 1999b] the soft-max function has been proposed as a biologically plausible implementation of the maximum.

Other important elements in the circuits are the monotonic nonlinearities, approximated here in terms of polynomial exponents p, q and r in the figure. They correspond to different strengths of nonlinear monotonic transfer functions due to synaptic and voltage-to-spike transduction. In this static approximation, sigmoid nonlinearities can be found in the cortex, for example, between presynaptic and postsynaptic potentials. Different operating ranges within a sigmoid transfer function can be approximated by polynomial functions with different exponents (p, q or r). Combined with divisive normalization, the relative strength of these nonlinearities can produce either selective or invariant responses, as illustrated in Table 2.1.

The most general form of output from these circuits is given by Eq. 2.3. If $p < qr$, y monotonically increases like $|\vec{x}|^p$ for small input, and for large input, y decreases and approaches 0 as the normalizing denominator grows faster than the numerator. In other words, the output y will show a tuning behavior, peaking at and being selective for a particular pattern of inputs. A simple calculation can show the relationship between the optimal input pattern, synaptic weights $\vec{w}$ and constant k for a given set of parameters p, q and r (see Appendix B.2).

An additional sigmoid-like nonlinearity may also operate on the output y:

$$h(y) = \frac{1}{1 + e^{-\alpha(y-\beta)}}. \tag{2.9}$$

The parameters α and β in Eq. 2.9 have the similar role as σ of the Gaussian function in Eq. 2.1, controlling the sharpness of tuning. These parameters may be learned and calibrated (e.g., large α and β to produce a sharp and sparse tuning), along with the synaptic weights, during the developmental or training period of the neural circuit.

A multiquadric radial basis function, $\sqrt{k^2 + |\vec{x} - \vec{w}|^2}$, which has an inverted Gaussian-like response profile, may be considered as another possible tuning operation [Girosi et al., 1995]. It can be computed with the same neural circuits A and B in Fig. 2-1, by the Gaussian-like tuning followed by an inverted sigmoidal nonlinearity (e.g., $1-1/(1+e^{-\alpha(y-\beta)})$). Although some neurons reportedly show monotonically increasing, multiquadric-like responses along some feature dimensions [Leopold et al., 2006; Freiwald et al., 2005], they may still be explained by the Gaussian-like computations operating in the neural input space, rather than the stimulus space.

A max-like output can be achieved by the same functional form with different parameters. Eq. 2.7 in Table 2.1 is the soft-max function, which is a widely used approximation to the maximum [Yu et al., 2002]. For small k and large q, it approaches $\max(\vec{x})$. Note that, unlike the Gaussian-like operation, the synaptic weights are $w_i = 1$ (for all i) and the strength of nonlinearity in the denominator is smaller than the numerator (i.e., $p = q+1 > q$). Thus, the same biophysical mechanism with different parameters (and potentially even in different operating regimes of the same circuit) can compute both Gaussian-like and max-like operations.

2.2.2 Gaussian-like Operation with L2-norm

An interesting insight can be gained by considering a particular form of normalization known as the L2-norm, which is the most commonly used vector normalization scheme: $|\vec{x}|_{L2} = \sqrt{\sum_{i=1}^{n} x_i^2}$. The L2-norm can be computed by the divisive normalization circuits A and B in Fig. 2-1 with $p = 1$, $q = 2$, and $r = 0.5$ [Heeger, 1993; Kouh and Poggio, 2004].

The key computation in a Gaussian function (Eq. 2.1) is $|\vec{x} - \vec{w}|^2$, which is a measure of similarity between two vectors and which generates a tuned response profile around an "optimal" pattern $\vec{w}$. A biologically plausible mechanism for such a computation is suggested by the following mathematical identity that relates the Euclidean distance measure, which appears in the Gaussian function, with the normalized scalar product:

$$|\vec{x} - \vec{w}|^2 = -2\vec{x} \cdot \vec{w} + 1 + |\vec{w}|^2, \text{if } |\vec{x}| = 1. \tag{2.10}$$

In other words, the similarity between two normalized vectors, $\vec{x}$ and $\vec{w}$, can be measured either by the Euclidean distance or by the scalar product (the angle between the two vectors in a multi-dimensional space). Hence, Eq. 2.10 suggests that a Gaussian-like tuning can arise from a normalized scalar product operation, as described in detail by [Maruyama et al., 1992].[2]

Note, however, that a Gaussian function may have a center at any arbitrary point in multidimensional input space, while a normalized scalar product, $\vec{w} \cdot \vec{x}/|\vec{x}|$, is tuned only to the direction of the vector, because the normalization condition reduces the dimensionality by one (also note that L2-norm does not satisfy $p < qr$ condition for the tuning, mentioned in the previous section, because $p = 1$, $q = 2$, and $r = 0.5$). Although it may not be necessary[3], an extra input can provide a simple solution to salvage the limitation of the L2-normalized scalar product operation (see Appendix B.4).

[2] In [Maruyama et al., 1992], the connections between a multilayer perceptron and a neural network with the radial basis functions are explored. Their analysis is based on the exact form of the identity in Eq. 2.10 (i.e., the input $\vec{x}$ to the Euclidean distance is normalized as well as the input to the scalar product). Here, a looser connection is examined, where the input to the Euclidean distance is not normalized, but the input to the scalar product is normalized:

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

[3] In a high dimensional input space, many features and tuning dimensions are available, as likely in the cortex. Therefore, a normalized scalar product (and the tuning to the direction of an input vector) may provide enough selectivity for an input pattern, so it may not be necessary to achieve a tuning to an arbitrary point in the input space like the Gaussian function.

2.2.3 Circuit C

In the limit of weak shunting inhibition, the general computation performed by these canonical circuits can still involve some normalizing components (see [Borst et al., 1995] and Appendix B.1). Under certain conditions (e.g., negligible normalization due to large leak current; see Appendix B.1), the operation may even reduce to a weighted summation, potentially subject to different monotonic nonlinearities at the synapses. The overall response of such a circuit increases with the increasing response by any of its afferents (or with the decreasing response by the afferent neurons with the negative synaptic weights). Therefore, the maximum response of the circuit is constrained to be the point with the maximum activations of the positive-weighted afferents and with the minimum activations of the negative-weighted afferents. Tuning to an arbitrary input pattern is not possible, unlike Eq. 2.3 or the Gaussian function.[4]

2.2.4 Comparisons

Fig. 2-2 shows in a two-dimensional input space that the proposed circuits in Fig. 2-1 can indeed approximate the Gaussian and maximum operations. Importantly, the approximations also work in higher dimensions (see Appendix B.6).

Circuits A and B in Fig. 2-1, while requiring more complex and precise synaptic connections, can provide better approximations to the reference functions (Eq. 2.1 and 2.2) and more intricate tuning behaviors than circuit C, which may be considered as a "poor man's" version of the canonical circuit. Circuit C is simpler, but it is expected to require a much larger number of afferent inputs to generate complex tuning behaviors. Hence, there is a trade-off between the complexity of tuning and the economics of maintaining multiple synapses. In fact, the actual mechanism in the cortex may be mixed or a hybrid of the two. Thorough and precise experiments at both intra- and extra-cellular levels will be necessary to identify the actual tuning mechanisms in the cortex.[5]

[4]When the number of the afferent neurons is large, however, the weighted summation operation may still create *almost* as complex a tuning behavior as the Gaussian function. The afferent response with an inhibitory synaptic weight (e.g., presence of an unpreferred feature in the stimulus) can lower the overall output, acting like a Gaussian function tuned to the small input. In other words, the inhibition effectively produces an *AND-NOT* behavior, and a combination of the *AND-NOT* operations can create a tuning behavior similar to the Gaussian function. The output of the weighted summation operation depends on the weights as well as the norm of the input vector. With a large input dimensionality, the norm of the input vector has less influence over the tuning behavior, as the fluctuations from each input component average out. In other words, the variance of the afferent response decreases with the increasing number of the afferent neurons (cf. $1/n$ dependence of the variance from the n-dimensional statistical samples). As a result, the weighted summation may yield a tuning behavior strongly modulated by the synaptic weights, although it still is a monotonic tuning within the input space of the afferent responses. An approximation to the maximum operation can also be computed by a weighted summation with monotonic nonlinearities: for example, $y = \log(\sum_{j=1}^{n} e^{x_j})$. Similar to the soft-max operation, the summation of strongly scaled input values (e.g., exponential function) followed by an inverse scaling operation will produce a max-like response. However, while the log-like nonlinearity is commonly found in the cortex, it is an open question whether the biophysics of a real neuron would have enough dynamic range to allow such a steep, exponential nonlinearity.

[5]There is indirect evidence that some cortical tuning operations are more like a Gaussian function rather than a simple weighted summation. Some experiments (e.g., [Sundberg et al., 2005] and personal communication with D. Zoccolan, 2006) have reported that increasing the contrast of a stimulus does not always saturate the overall output of a neuron. Instead, in some cases it may decrease the response. Assuming that higher stimulus contrast typically increases the afferent responses, the decreasing output might indicate that the neuron is tuned to an intermediate level of the afferent activations. Another possible explanation is that a pool of inhibitory neurons is turned on or overtakes the excitatory inputs only at higher stimulus contrast, yielding a suppressive effect.

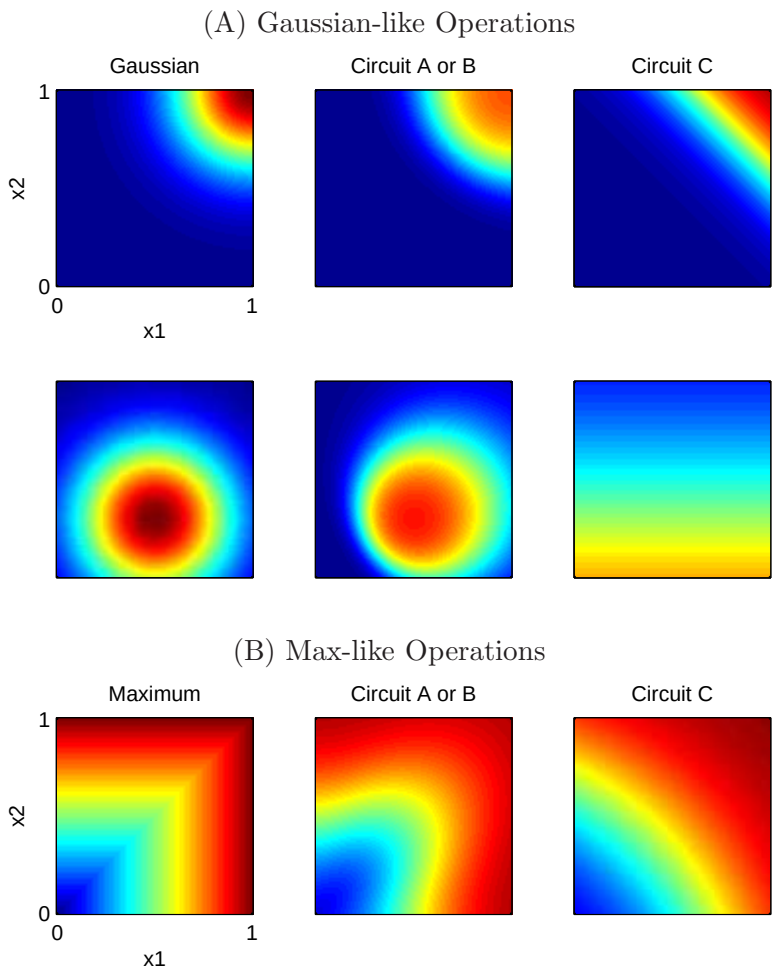


Figure 2-2: This figure illustrates that Eq. 2.3 in Table 2.1, possibly generated by the circuits in Fig. 2-1, can approximate the Gaussian and maximum operations in a two-dimensional input space. Within each panel, the x_1 and x_2 axes represent the responses of two afferent neurons, bounded between 0 and 1, where 1 is the maximum response. The output of the efferent neuron is represented by the color scale. (A) The Gaussian functions (first column) centered at $(1, 1)$ and $(0.5, 0.3)$ are approximated by the divisive normalization operation, Eq. 2.3 with $(p, q, r) = (1, 2, 1)$ (second column: circuit A or B) and the weighted summation, Eq. 2.4 (third column: circuit C). The sigmoid nonlinearity on the output, Eq. 2.9 and the synaptic weights for Eq. 2.4 are found by a numerical fitting routine. Note that circuits A and B can be selective for an arbitrary point in the input space, while circuit C can not. Although not as flexible as circuit A or B, circuit C may exhibit more complex behaviors (like an *AND-NOT* operation) in a higher dimensional input space, so the qualitative differences between these circuits may be smaller. (B) The maximum operation (first column) can also be approximated by the soft-max operation, Eq. 2.3 with $(p, q, r) = (3, 2, 1)$ (second column). The sigmoid nonlinearity has been also applied on the output. For Eq. 2.4 (third column), the sigmoid parameters and the synaptic weights are again fitted numerically. In general, the divisive normalization operations provide better approximations to the Gaussian and the maximum operations than a simple weighted summation. See Appendix B.6 for the approximations in higher input dimensions.

2.2.5 Learning the Synaptic Weights

Now given the fact that both Gaussian-like and max-like operations may be implemented by relatively simple, biologically plausible neural circuits, it is then an interesting question whether these operations can be learned also from a similarly simple and plausible plasticity mechanism, especially for setting the values of the synaptic weights within the neural circuits in Fig. 2-1. According to Table 2.1, the synaptic weights implementing the max-like operation are all uniform ($w_i = 1$ for all i), indicating that the circuit for the max-like operation probably requires less fine-tuning of the synaptic weights. On the other hand, the Gaussian-like tuning operation will likely require more precise tuning of the synaptic weights [Serre et al., 2005].

Hebb's rule ("fire together, wire together") is the best known synaptic learning rule in neuroscience. Consider the following modified Hebbian rule with a flavor of the stochastic gradient descent algorithm ($\Delta w \propto \Delta y$):

$$w(t + \Delta t) = w(t) + \eta \cdot \big(y(x, t + \Delta t; w + \eta) - y(x, t; w)\big), \tag{2.11}$$

where η is a small random jitter in a synaptic weight. The synaptic weight is strengthened if this random perturbation produces a higher response, and it is weakened if the perturbation produces a lower response. Thus, the "wiring" of the synapse is dependent on not just whether the presynaptic and postsynaptic neurons fire together, but also whether the presynaptic activity has a positive effect on the postsynaptic neuron within a short temporal window [Dan and Poo, 2004] (hence, *stochastic gradient descent*).[6]

Now suppose that a neural circuit A or B in Fig. 2-1 is repeatedly exposed to a particular pattern of input $\vec{x}_o$, so that a target value for $\vec{w}$ is given. Then, the learning rule, Eq. 2.11, brings the synaptic weights to the target after several iterations (i.e., $\vec{w} \to \vec{w}_o$, so that $y(\vec{x}_o; \vec{w}_o) =$ maximum output value), as shown by the simulation results in Fig. 2-3.

This scheme involves supervised learning, because the input $\vec{x}$ is required to be fixed at some $\vec{x}_o$ (e.g., the system is exposed to a target stimulus) while the learning takes place. However, choosing the target values themselves may be unsupervised or subject to a hedonistic principle [Seung, 2003]. As demonstrated in [Serre et al., 2005], the selectivity for the features useful for object recognition (i.e., the centers of the Gaussian-like tuning functions) can be learned by storing the snapshots of the naturally occurring sensory stimuli

[6] If the tuning operation does not have a divisive normalization or a Gaussian-like response profile, an additional constraint is required for stability. One such example is the following:

$$\sum_{i=1}^{n} w_i^2(t) = w_o. \tag{2.12}$$

This condition, also known as Oja's rule [Oja, 1982], implies that the total synaptic weights are normalized or conserved. When some synaptic weights are increased, others will be decreased. This particular mathematical form (i.e., squaring of the weights) is chosen for convenience, but captures the basic idea. The normalization of the synaptic weights is a separate mechanism, different from the normalization of the inputs. Such a condition is expected since the synapses are likely competing for the limited resources (e.g., the number of the receptors on the postsynaptic membrane). Without any mechanism for competition, a simple Hebb-like rule will be subject to a positive feedback (stronger synapses getting even stronger), leading to instability. There is a growing body of experimental evidence and theoretical arguments in support of the competitive Hebbian mechanism [Miller, 1996]. Alternatively, a decay term in the learning rule can also ensure stability, as done in [Földiák, 1991]. In the case of the normalization circuit where the response does not always increase with the larger norm of $\vec{w}$, Eq. 2.11 converges to a stable value without an extra constraint like Eq. 2.12.

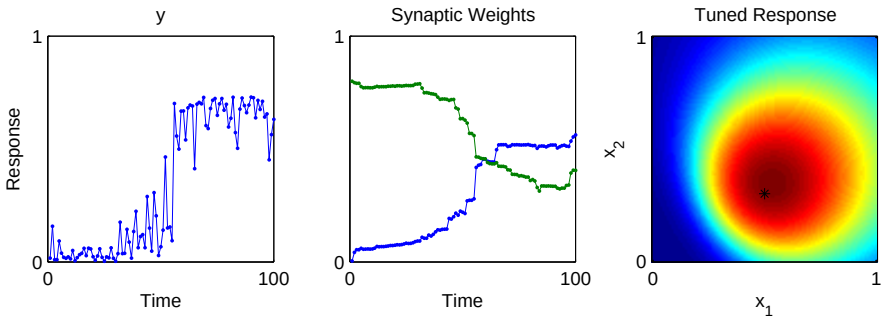


Figure 2-3: The neural circuit performing the Gaussian-like operation can be tuned to a particular input pattern, by learning the synaptic weights. The example shown here has two afferent inputs (x_1 and x_2), and the target pattern has been set at $\vec{x}_o = (0.5, 0.3)$. While the presynaptic activations are fixed at $\vec{x}_o$, the synaptic weights evolve according to Eq. 2.11 from randomly selected initial values (middle figure). The left figure shows the evolution of the output of Eq. 2.3 (with $(p, q, r) = (1, 2, 1)$ and a sigmoid on y), which reaches the maximum as the correct synaptic weights are learned. The right figure shows the learned tuning behavior around $\vec{x}_o$ marked by *. The random jitters, corresponding to η in Eq. 2.11, were selected from a normal distribution with the maximum value of 0.1. See Appendix B.6 for the examples of learning in higher input dimensions.

in an unsupervised manner. Such a learning scheme can provide a large, overcomplete set of features or basis functions that reflect the stimulus statistics inherent in the natural environment. Also note that the invariance property can be learned by the similar Hebbian mechanism [Földiák, 1991; Bartlett and Sejnowski, 1998].

2.3 Discussion

In summary, the same neural circuit can perform both Gaussian-like and max-like operations. There are other models of the Gaussian-like tuning and max-like operations. For example, "pattern-matching" may occur in the dendritic spines [Blackwell et al., 1998] or in a laminarly organized neural circuit [Douglas et al., 1989]. Translation invariance may be computed by the intra-dendritic mechanisms [Mel et al., 1998] or by the winner-take-all types of neural circuitry [Rousselet et al., 2003; Hahnloser et al., 2000; Yuille and Grzywacz, 1989]. Here, a unifying model of these two neural operations was proposed, with a postulate that one of the main roles for the wide-spread gain control found in the cortex is to implement Gaussian-like and max-like operations.

As shown by several experimental studies [Wang et al., 2000; Marino et al., 2005; Douglas et al., 1989], inhibitory mechanisms are critically involved in generating neural selectivity and invariance, and in the current proposal, the input normalization by a pool cell or lateral inhibition (Fig. 2-1) is the major component in the neural circuit. The inhibitory pool cells or interneurons receive input from the afferent neurons with different receptive fields and selectivities. As a result, these inhibitory cells are expected to have a large receptive field (i.e., the union of the receptive fields of the afferent neurons), broad selectivity and typically high responses. Interestingly, putative inhibitory neurons possess similar traits [Bruno and Simons, 2002; Frank et al., 2001].

Comparing the circuits A, B, and C in Fig. 2-1, it was noted that there is a trade-off

between the complexity of the circuit and the number of the afferent inputs (i.e., more complex circuit can generate more elaborate tuning behaviors with fewer inputs). A hybrid strategy may be employed in the cortex, where the initial selectivity of a neuron may be determined initially by the set of the afferent neurons, and the finer, more complex selectivity may develop through the secondary inhibitory mechanisms (e.g., through divisive normalization). Such a refinement of selectivity can be distinguished from the response modulations from the feedback or attentional mechanisms (probably operating via back-projecting connections in the hierarchy of the cortical areas), because it will have a short latency (in the order of ten milliseconds, assuming only a minor short delay in the pooling stage). There are some experimental results showing the dynamics of the neural selectivity (in V1 [Shapley et al., 2003], in V2 [Hegde and Van Essen, 2004] and in IT [Brincat and Connor, 2006]), but their mechanisms and relevances to the current proposal are not clear.

The idea that the neural selectivity ultimately arises from a combination of the feedforward inputs and the intra-cortical refinements takes an intermediate position between the two competing theories about the origin of orientation selectivity in the primary visual cortex V1 [Ferster and Miller, 2000]. That is, the overall orientation preference of a V1 neuron is initially determined by the spatial arrangements of the LGN afferent inputs, and then the selectivity is sharpened by lateral interactions. Furthermore, note that the model of motion selectivity in MT by [Rust et al., 2006] is equivalent to the proposed neural circuitry in Fig. 2-1, since they both involve weighted summation over the divisively normalized inputs, despite some differences (e.g., the model of Rust et al. allows for the negative synaptic weights and for the stronger self-normalizing terms). It is quite interesting that similar neural circuits may be operating not just in the ventral pathway (as in Fig. 3-1), but also in the dorsal pathway.

The requirement for balanced selectivity and invariance is a universal computational principle for robust object recognition, and it is applicable for all sensory modalities. Moreover, there is a similarity in the putative learning schemes for selectivity and invariance. Acquiring invariance involves learning correlations within temporal sequences of object transformations, as proposed in [Földiák, 1991; Bartlett and Sejnowski, 1998; Wallis and Bulthoff, 1999]. Similarly, acquiring selectivity for a particular object involves learning correlations over features (e.g., particular arrangements of the parts and the features of a stimulus) [Serre et al., 2005].

Along with the universality of these computational and learning principles, the universality of the proposed neural circuit is consistent with the interesting observation that the basic cortical structure is quite similar within and across different functional areas and that there may exist "canonical microcircuits" in the cortex [Mountcastle, 2003; Nelson, 2002; Douglas et al., 1989]. It has even been proposed that the functional distinction between the simple and complex cells in the primary visual cortex should be replaced by a continuous spectrum within a neural population [Chance et al., 1999; Mechler and Ringach, 2002; Priebe et al., 2004]. Many of the neural circuit models in the literature indeed have similar operating principles that are different only in their details (e.g., the excitatory and inhibitory circuit elements may be found at specific laminar locations within cortical tissues [Douglas et al., 1989], or the inhibition may operate through presynaptic feedback [Yuille and Grzywacz, 1989], instead of a feedforward pooling mechanism as in Fig. 2-1A). Clearly, the study of more detailed, biophysical implementations of the proposed static neural circuits (e.g., spiking neural network including the kinetics of realistic synapses and of the spike generation process; see Section 5 in [Serre et al., 2005] and [Knoblich et al., 2007]) is the next important step in the work described here.

Chapter 3

Model Overview

In the previous chapter, it was proposed and demonstrated that the same neural circuitry can perform both Gaussian-like and max-like operations. This chapter introduces a computational model of a part of the primate visual cortex, namely the ventral pathway. This model employs these canonical neural operations for explaining visual object recognition, mediated by the neurons along the ventral pathway. Such a model provides a unifying framework for quantitatively understanding different sets of experimental data and making a coherent picture and testable predictions.

The original implementation of the model, formerly known as HMAX, is described in [Riesenhuber and Poggio, 1999b]. The current, extended implementation is described in [Serre et al., 2005], covering in detail many other interesting aspects that are beyond the scope of this thesis.

3.1 Object Recognition and a Model of the Ventral Pathway

Object recognition is one of the main functions that the cortex can perform with phenomenal speed and efficiency. For obvious reasons, the ability to recognize an object is important for an organism's survival (e.g., distinguishing a prey from a predator), and all sensory modalities (vision, audition, olfaction, etc.) are involved in this task, sometimes independently and other times cooperatively.

While done seemingly effortlessly by the cortex, object recognition is a difficult computational problem because there are two conflicting requirements. On one hand, the recognition system needs to be selective for different objects that can in general be similar and share certain characteristic features (e.g., telling apart identical twins). On the other hand, the recognition needs to be invariant (or tolerant) to various appearance-altering transformations of an object (e.g., recognizing a person under disguise).

The picture emerging from decades of physiology, psychophysics, and neuroscience research is that the cortex solves such a difficult computational problem by gradually increasing and building up selectivity and invariance with a hierarchical architecture [Hubel and Wiesel, 1962; Fukushima et al., 1983; Perrett and Oram, 1993; Wallis and Rolls, 1997; Riesenhuber and Poggio, 1999b; Mel, 1997]. Note that most of these models, including the one in Fig. 3-1, only focus on the feedforward operations and disregard the effects of feedback or attention. Although the recognition performance in demanding tasks can be improved significantly by the attention and longer exposures to the stimuli, many experiments have suggested (for example, by presenting a rapid sequence of visual images [Thorpe

et al., 1996; Hung et al., 2005]) that humans and primates can correctly recognize objects even when the feedback, recurrent or attentional processing is blocked (e.g., by masking) or does not have enough time or resources to be invoked [Li et al., 2002]. However, a feedforward operation may involve some *local* feedback or lateral connections (different from the major feedback or back-projections mentioned above), as long as its neural circuit employs a forward directional flow of computations and processing. Such a computational strategy has been proposed as a way of effectively dealing with the "binding problem" [Mel and Fiser, 2000; Riesenhuber and Poggio, 1999a] and satisfying the conflicting requirements for the selectivity and invariance.

A computational model [Serre et al., 2005; Riesenhuber and Poggio, 1999b], as shown in Fig. 3-1, is constructed to closely follow the anatomy and physiology of the primate visual cortex. A Gaussian-like and a max-like operations are repeatedly applied, progressively building up richer and more complex selectivity and invariance properties along the hierarchical framework. The selectivity of the model units in the intermediate layers are learned in an initial developmental-like unsupervised learning stage, possibly using a mechanism like Eq. 2.11, so that they respond robustly to the features found in the stimulus set or in natural images.[1]

The current version of the model [Serre et al., 2005] is an extension of the original formulation [Riesenhuber and Poggio, 1999b] in several ways: (1) The optimal activation patterns for S units, performing the Gaussian-like tuning operations, are more varied to account for the diverse selectivity properties measured in the cortex. Such diverse tuning properties are learned in an unsupervised way, by using thousands of natural images during a developmental stage; (2) The Gaussian-like and max-like operations have more biologically plausible forms, Eq. 2.3; (3) The new version of the model better matches the anatomy and physiology of the ventral pathway, by the introduction of additional layers (more closely comparable to the visual areas V4, posterior and anterior IT) and anatomically-observed bypass routes. These changes were natural and planned extensions of the original model, and further information can be found in [Serre et al., 2005].

In summary, the overall architecture of the model is hierarchical and feedforward, reflecting

1. the hierarchical organization of the ventral pathway [Felleman and Van Essen, 1991],

2. the rapid recognition performance that does not leave much time for feedback or recurrent processes [Thorpe and Imbert, 1989; Thorpe et al., 1996; Hung et al., 2005] (hence, focusing on the first few hundred milliseconds of visual processing without attention),

3. and the gradual build-up of more complex selectivity and invariance along the hierarchy [Kobatake and Tanaka, 1994] (by repeatedly employing Gaussian-like and max-like operations in an interleaved fashion throughout the hierarchy).

The overall architecture of the model is implemented with three components: Neurons, synaptic weights, and operations. Each model unit (a "neuron") is connected to a number of afferent or input units (i.e., the "presynaptic neurons"). Either Gaussian-like or max-like operation (depending on the functional role of each model unit) is performed over the

[1] Learning in the model is based on a very simple procedure of "taking snapshots" of the training images [Serre et al., 2005]. That is, an activation pattern due to a particular patch of a stimulus image is stored by synaptic weights. Hence, the neuron will activate most strongly when the same image is encountered again, and will respond less to other images.

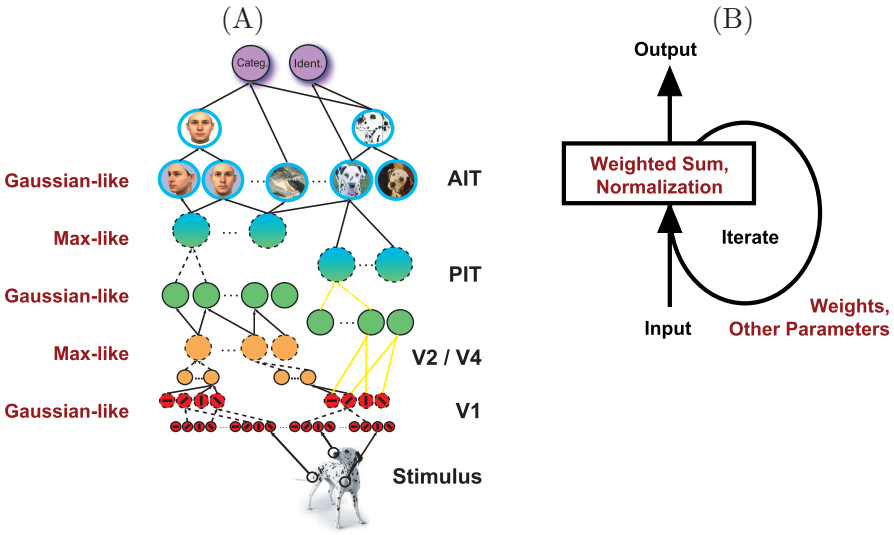


Figure 3-1: (A) The architecture of a model of the ventral pathway in the primate visual cortex for object recognition. The model describes the early stages of the visual information processing for "immediate" recognition. Gaussian-like (solid lines) and max-like (dashed lines) operations are repeatedly applied along the hierarchy in an interleaved manner. As a result of such construction, progressively more tolerant (to changes in position and scale) and more complex neural selectivities are generated, as observed along the ventral pathway. See [Serre et al., 2005] for a comprehensive description and [Riesenhuber and Poggio, 1999b] for the original proposal. (B) In implementing this model, the same neural operation Eq. 2.3 (i.e., the weighted sum and divisive normalization with different parameters, possibly computed by the canonical neural circuits) has been used to achieve both selectivity and invariance. During each iteration or at each layer, the synaptic weights and other parameters in the circuit are specified. The software of the model is available at `http://cbcl.mit.edu`.

afferent responses (a continuous value between 0 and 1, in this case), using a weighted summation (with synapses) and a divisive normalization (with inhibitory connections), which can be implemented by a canonical neural circuitry. Given the hierarchical and feedforward nature of the model, these three components determine the output of the model to any stimuli. See Appendix A.1.

3.2 Correspondence with Anatomy and Physiology

The model is constructed to reflect the anatomy and physiology of the ventral pathway, so that the hierarchical layers in the model correspond to the visual areas V1, V2, V4, and IT [Felleman and Van Essen, 1991], and the receptive field sizes and tuning properties of the model units closely match those of the neurons in the corresponding areas.

The ventral pathway is one of two major cortical pathways that process visual information, and it has been closely linked to object recognition by a variety of experiments (for a review, see [Ungerleider and Haxby, 1994]). Many studies have explored and described the representations at various stages along the ventral pathway. These studies have shown that the responses of neurons in lower visual areas, such as primary visual cortex (V1), and higher visual areas, such as inferior temporal (IT) complex, explicitly represent features or information about visual form. Neurons in V1 have small receptive fields and are responsive to simple features, such as edge orientation [De Valois et al., 1982; Hubel and Wiesel, 1962], while neurons far along the pathway in IT have large receptive fields and can be selective for complex shapes like faces, hands, and specific views of other familiar objects [Gross et al., 1972; Hung et al., 2005; Logothetis et al., 1995; Tanaka et al., 1991].

Neural response properties in areas V2 and V4, which lie between V1 and IT, reflect their intermediate anatomical position. Their receptive field sizes are, on average, greater than those in V1, but smaller than those in IT [Desimone and Schein, 1987; Kobatake and Tanaka, 1994]. Many V4 neurons are sensitive to stimulus features of moderate complexity [Desimone and Schein, 1987; Freiwald et al., 2004; Gallant et al., 1996; Gawne and Martin, 2002; Kobatake and Tanaka, 1994; Pasupathy and Connor, 1999, 2001; Pollen et al., 2002].

In the lowest layer of the model (corresponding to V1 simple cells), the model units have Gabor receptive field profiles at multiple sizes and orientations. In the next layer (corresponding to V1 complex cells), the softmax operation (Eq. 2.7) is performed over the units in the previous layer with identical selectivity but with slightly shifted and/or scaled receptive fields, similar to the classical model of V1 by Hubel and Wiesel (1962, 1965). The parameters for these layers have been chosen to reflect several experimental findings about the properties of the simple and complex cells in V1 [Serre et al., 2005].

The same sequence is repeated in the next two layers, where the Gaussian-like tuning operation and the softmax operation are performed on their respective afferent units. These layers loosely correspond to visual areas V2, V4, and PIT, based on their receptive field sizes and their tuning properties to shapes and features of intermediate complexity. Instead of being hard-wired (unlike the Gabor-like orientation selectivity in the lowest layer), the selectivity of the units in the intermediate layers are learned in an unsupervised way by being exposed to different images, during a "developmental" stage.

The next stage corresponds to AIT, where neurons with very complex shape selectivity, large receptive field, and highly invariant response properties are found. Beyond this stage, a task-dependent, decision-making circuit, corresponding to PFC, is also assumed to receive feedforward inputs from the lower areas.

Chapter 4

Comparison with the V4 Neurons

Neurons in macaque monkey area V4, an intermediate stage along the ventral pathway, have been shown to exhibit selectivity for various shapes and features of intermediate complexity (e.g., non-Cartesian gratings, stimuli containing multiple orientations, boundary conformations, etc.) and invariance to spatial translation [Kobatake and Tanaka, 1994; Desimone and Schein, 1987; Gallant et al., 1996; Pasupathy and Connor, 2001].

In this chapter, it is demonstrated that the model units in the intermediate layer, corresponding to V4, can successfully generate similar shape selectivity and invariance properties through a nonlinear, translation-invariant combination of locally selective subunits, suggesting that a similar transformation may occur or culminate in area V4. More detailed version of this work can be found in [Cadieu et al., 2007]. The development of the V4 model and its analysis was done in close collaboration with Charles Cadieu (now at UC Berkeley).

4.1 Physiology of Area V4 and the Model

4.1.1 V4

Area V4 lies in the middle of the ventral pathway, which is one of two major cortical pathways that process visual information and which has been closely linked to object recognition by a variety of experiments (for a review see [Ungerleider and Haxby, 1994]). A lesion in this area results in the impairment of shape perception and attention [De Weerd et al., 1996; Gallant et al., 2000; Girard et al., 2002; Merigan and Pham, 1998; Schiller, 1995; Schiller and Lee, 1991], suggesting that V4 is likely to play a critical role in object recognition. Neural response properties in area V4 reflect its intermediate anatomical position. V4 receptive field sizes average 4-7 times those in V1, but are smaller than those in IT [Desimone and Schein, 1987; Kobatake and Tanaka, 1994]. Many V4 neurons are sensitive to stimulus features of moderate complexity and, at the same time, invariant to local translations [Desimone and Schein, 1987; Freiwald et al., 2004; Gallant et al., 1996; Gawne and Martin, 2002; Kobatake and Tanaka, 1994; Pasupathy and Connor, 1999, 2001; Pollen et al., 2002].

Previously, Pasupathy and Connor (1999, 2001) provided a quantitative, phenomenological description of stimulus shape selectivity and position invariance in area V4. They demonstrated that a subpopulation of V4 neurons, screened for their high firing rates to complex stimuli, is sensitive for local modulations of boundary shape and orientation. The responses of these neurons can be described as basis function-like tuning for curvature, orientation, and object-relative position of boundary fragments within larger, more complex global shapes. This tuning is relatively invariant to local translation. At the population

level, a global shape may be represented in terms of its constituent boundary fragments by multiple peaks in the population response pattern [Pasupathy and Connor, 2002]. Brincat and Connor showed that V4 signals for local boundary fragments may be integrated into more complex shape constructs at subsequent processing stages in posterior IT [Brincat and Connor, 2004, 2006].

4.1.2 Model of V4

Within the full model of the ventral pathway, Fig. 3-1 [Serre et al., 2005], the C2 units in the intermediate layer represent V4 neurons. The lower S1, C1, and S2 units of the model are analogous to neurons in the visual areas V1 and V2, which precede V4 in the feedforward hierarchy (the role of V2 and the issue of anatomical correspondence for the S2 layer are considered later in the Discussion section). The key parts of the resulting V4 model are summarized schematically in Fig. 4-1.

This V4 model is consistent with several other quantitative and qualitative models of V4 (e.g., [Gallant et al., 1996; Li, 1998; Reynolds et al., 1999; Wilson and Wilkinson, 1998]), where several orientation-selective afferent units are combined with nonlinear feedforward operations, often involving inhibitory elements. The divisive normalization operation used in the model (Eq. 2.3) is closely related to, for example, the center-surround inhibition in Wilson and Wilkinson (1998) and especially the biased-competition model formulation in Reynolds et al. (1999). Those models have been successful in describing and explaining different specific phenomena, such as texture discrimination [Wilson and Wilkinson, 1998], contour integration [Li, 1998], or attentional effects [Reynolds et al., 1999], occurring in or around V4.

The model presented in Fig. 4-1, however, differs from others in the following aspects. First, the role of area V4 is considered within a larger framework that attempts to explain the entire ventral pathway at a computational and quantitative level. Second, the model not only attempts to explain experimental findings, but also attempts to explain how V4 responses could be computed from the known properties of earlier stages within the ventral pathway. Third, the model involves two stages of computation to account for the selectivity of V4 neurons to complex stimuli and their invariance to visual translations.

In the feedforward direction, the gray-level pixel values of a stimulus image are processed first by the S1 units that correspond to "simple" cells in V1. They have Gabor receptive field profiles with different sizes and four orientations (0, 45, 90, and 135 degrees). The responses are determined by the normalized dot product of Eq. 2.3 with $(p, q, r) = (1, 2, 1/2)$ and $k = 0$. The sigmoid nonlinearity is not used here for simplicity. This results in a model of simple V1 neurons similar to that presented in [Carandini and Heeger, 1994; Heeger, 1993]. S1 responses were rectified by taking their absolute value, equivalent to having rectified S1 units of opposite signs (e.g., on-off and off-on cells) project to the same efferent units.

C1 units, which correspond to "complex" V1 cells, perform the invariance operation (maximum) over S1 units with identical orientation selectivity, but slighted shifted or scaled receptive fields. As a result of such construction, C1 units have orientation selectivity with larger receptive fields than S1 units, within which translation and scale invariance is observed, similar to complex V1 cells. Three different spatial pooling ranges over S1 units are used to created C1 units with varying receptive field sizes. The receptive fields of adjacent C1 units (with the same size) overlap by half.

The same construction principle (alternating between Gaussian-like tuning and maximum-like operations) is repeated in the next two layers, S2 and C2. S2 units perform the nor-

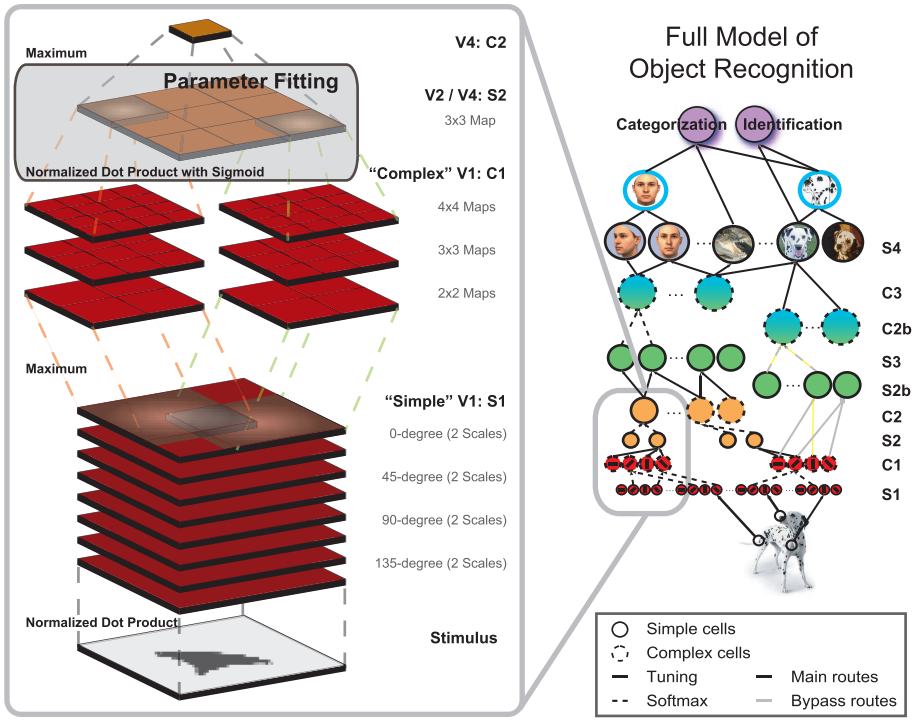


Figure 4-1: The model of V4, shown on the left, is part of an extensive theory of object recognition, shown on the right, dealing with the computations and neuronal circuits in the feedforward pathway of the ventral stream in primate visual cortex [Riesenhuber and Poggio, 1999b; Serre et al., 2005]. The response of a C2 unit (at the top of the left diagram) is used to model the responses of individual V4 neurons and is determined by the preceding layers of units, corresponding to earlier stages of the ventral pathway before area V4. The buildup of selectivity and invariance is implemented by the alternating hierarchical layers of "simple" S units, performing Gaussian-like tuning operations (by normalized dot product), and "complex" C units performing maximum operations. The designation of "simple" and "complex" follows the convention of distinguishing between the orientation-tuned, phase-dependent simple cells and the translation-invariant complex cells of V1 [Hubel and Wiesel, 1962, 1968]. Because V4 neurons exhibit both selectivity for complex shapes and invariance to local translation, V4 neurons are modeled with the responses of C2 units by the combination of translated copies of S2 unit afferents with identical selectivity, but shifted receptive fields, following the same construction principle as in S1 and C1 layers. The lower S1 and C1 units of the model are analogous to neurons in area V1. In the feedforward direction, the image is processed by "simple" V1-like S1 units which send efferent projections to "complex" V1-like C1 units (for clarity, only a subset of C1 units are shown). S2 units receive input from a specific combination of C1 unit afferents and are selective for a particular activation of those inputs. Finally, the C2 unit pools over shifted S2 units. The resulting C2 unit produces a high response to a specific stimulus and is invariant to the exact position of the stimulus within the receptive field (the full receptive field spans the union of the receptive fields of S1 units). For different, non-optimal stimuli, the C2 response falls off as the afferent activity deviates from the optimal pattern. Most parameters in the model are fixed, except for the C1 and S2 connectivity (indicated by shaded rectangular region), which is varied to fit the individual neural responses.

malized dot product, Eq. 2.3 with $(p, q, r) = (1, 2, 1/2)$, on their C1 afferents, generating selectivity for features or shapes more complex than just orientation selectivity. Within the receptive field of each S2 unit, there are C1 units with three different receptive field sizes. The C1 units with the smallest receptive field size span the S2 receptive field in a 4x4 array, while the C1 units with larger receptive field sizes spanned the same S2 receptive field in a 3x3 or 2x2 array. Therefore, within each S2 receptive field, there are 29 (2x2+3x3+4x4) spatially locations, each with C1 units at four different orientations, resulting in a total of 116 (29x4) potential C1 units that could provide an input to an S2 unit. A small subset of these 116 C1 units is connected to an S2 unit, and the specific connectivity for each S2 unit (i.e., synaptic weights), along with the sigmoid parameters (Eq. 2.9) on the output of the S2 unit, can be fitted to a given V4 neuron's response profile.

The top level C2 unit, which corresponds to a V4 neuron, performs the maximum operation on the afferent projections from the S2 layer. Because V4 neurons exhibit both selectivity for complex shapes and invariance to local translation, V4 neurons are likely to combine translated copies of inputs with the same, but shifted, selectivity, just like the construction of a V1 complex cell. According to experimental studies [Desimone and Schein, 1987; Gallant et al., 1996; Pasupathy and Connor, 1999, 2001], V4 neurons maintain selectivity to translations of about 0.5 times the classical receptive field size. To match these experimental findings, a C2 unit received input from a 3x3 spatial grid of S2 units with identical selectivity properties, each shifted by 0.25 times the S2 receptive field (i.e., one C2 unit receives inputs from 9 S2 units). As a result, the C2 unit adopts the selectivity of its afferent S2 units to a particular pattern evoked by a stimulus in C1 and is invariant to the exact position of the stimulus. The C2 parameters, controlling the receptive field size and the range of translation invariance, are fixed throughout all the simulations.

In summary, this model of V4 is composed of hierarchical layers of model units performing feedforward selectivity or invariance operations. Most of the parameters are fixed to reasonable estimates based on experimental data from areas V1 and V4. To model a particular V4 neuron, only the parameters governing the connectivity between C1 and S2 layers, as indicated by the shaded rectangular region in Fig. 4-1, are found according to the fitting technique described below.

4.1.3 Fitting the V4 Responses

Using the model of V4, the electrophysiological responses of 109 V4 neurons previously reported in Pasupathy and Connor (2001) are examined, using the same set of parameterized contour feature stimuli, which have been constructed by a partial factorial cross of curvature values (sharp to shallow convex and concave curvature) at 45°-interval angular positions, relative to object center. The response of V4 neurons in awake Macaque monkeys to each stimulus shape during a 500 ms presentation period was averaged across 3-5 repetitions. For the analyses presented here, each neuron's responses across the entire stimulus set were normalized to range between 0 and 1.

Potentially, a number of different parameters in the model can be adjusted to match the selectivity and invariance profiles of the V4 responses. However, the selectivity of a C2 unit, which corresponds to a V4 neuron, is most dependent on the spatial arrangement and synaptic weights connecting the C1 units to the S2 units (modifying other parameters had little effect on the level of fit, see [Cadieu et al., 2007]). Furthermore, the model layers before S2 were not adjusted because they are considered analogous to representations in V1 and were not the focus of this study. The invariance operation at the C2 layer

was not adjusted because experimental results indicate that translation invariance over measured V4 populations is highly consistent [Desimone and Schein, 1987; Gallant et al., 1996; Pasupathy and Connor, 1999, 2001] and because the experimental measurements modeled here do not include sufficient stimuli at different translations. Therefore, the fitting algorithm determined the parameters of the selectivity operation at the S2 layer while holding all other parameters fixed (the fitted parameters within the overall model are indicated by the shaded box in the left panel of Fig. 4-1 labeled as "Parameter Fitting"). These fitted parameters include the subset of C1 afferents connected to an S2 unit, the connection weights to those C1 afferents, and the parameters of the sigmoid function that nonlinearly scaled the response values. For a given C2 unit, the parameters for all 3x3 afferent S2 units were identical to produce identical tuning over translation.

Because the model's hierarchy of nonlinear operations makes analytical solutions intractable, a greedy search algorithm was employed for its simplicity and efficacy in this problem domain. Fig. 4-2 shows an overview schematic of the fitting procedure and its implementation is described in detail in [Cadieu et al., 2007]. In short, this algorithm finds the best combination of C1 subunits and the best sigmoid parameters of the S2 unit by incrementally adding C1 afferents and minimizing the mean squared error between the experimentally measured V4 response and C2 unit's response.

Depending on the type of analysis, the number of C1 subunits of the best fitting model is determined by one of two methods. The first method was used to find a single model for each V4 neuron (as in Fig. 4-3 and 4-4) and used cross-validation to mitigate overfitting. In this method, the number of subunits was set to the minimum number of units, between 2 and 25, that minimized the average testing error over a 6-fold cross-validation set to within 1% of the absolute minimum. An n-fold cross-validation divides the dataset into n equally sized and randomly selected subsets, trains the model on $n-1$ of the subsets, and predicts the response on the remaining subset. This is repeated n times, each time predicting a different subset. Subsequently, the best fitting C2 unit with this number of subunits was found over the entire dataset. For each C2 unit, the maximum number of subunits was 25.

The second method was used to test the model's ability to generalize to stimuli outside the training set. The stimulus set was again splitted into randomly selected training and testing sets containing 305 and 61 stimulus-response pairs, respectively. The number of C1 subunits was determined on the training set by adding subunits in the greedy search until the error between the C2 unit's response and the V4 neuron's response decreased by less than 1% or once 25 C1 subunits were found. The resulting C2 unit's response was simulated on the test set, measuring the model's ability to generalize to stimuli outside the training set (as in Fig. 4-5).

In summary, the model of a V4 cell is derived from a parameter space consisting of 119 free parameters (synaptic weights for 116 C1 units and 3 sigmoid parameters: α, β, and the amplitude of the sigmoid function). The sigmoid parameters, α and β, control the steepness of tuning, and a numerical factor, multiplying the sigmoid function of Eq. 2.9, represents the maximum response of the S2 unit. A small number k (0.0001) prevents division by zero. For fitting the responses of each V4 neuron over 366 stimuli, a small subset of these parameters is selected based on cross-validation criteria.

One of the main limitations of this fitting framework is the stability of the solution. In other words, for a given response profile of a V4 neuron, the geometric configuration of the C1 subunits, obtained by the fitting procedure, is under-constrained and not guaranteed to be unique because there exist other configurations that would yield a similar level of fit with the neural response. However, most fitting results converged onto similar geometric

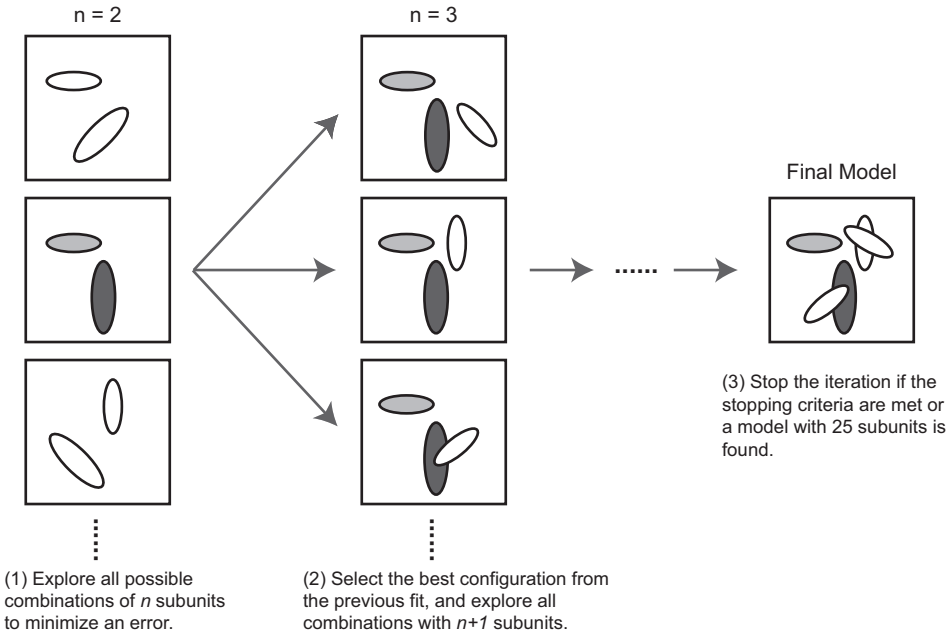


Figure 4-2: Schematic of the Model Fitting Procedure. The response of each V4 neuron was fit with a model C2 unit by determining the parameters between the C1 layer and the S2 layer (see "Parameter Fitting" box in Fig. 4-1). Because the response of C2 units is highly non-linear and an analytic solution is intractable, numerical method was used to find solutions, namely the set of C1 subunits connected to the S2 units, the weights of the connections, and the parameters of the sigmoid function. To determine the subset of C1 subunits, a forward selection algorithm, greedy search, was employed to find a solution. The search was initialized, column 1, by selecting two C1 subunits to include in the selectivity function. For each selection of two subunits, the parameters of the selectivity function were adjusted using gradient descent in parameter space to minimize the mean squared error between the V4 neuron's measured response and the model C2 unit's response. The C1 subunit configuration that achieved the lowest mean squared error was then used for the next iteration of the greedy search. The search then continued, column 2, by adding an additional C1 subunit to the best configuration found in the previous iteration. The search was stopped, column 3, to produce a final model. One of two stopping criterion were chosen based on the desired analysis of the final model. To find a single model for each V4 neuron, the search was halted once the average testing error over a 6-fold cross-validation set reached within 1% of the absolute minimum. To test the model's ability to generalize to stimuli outside the training set, the search was stopped once the mean squared error on the training set decreased by less than 1% or once 25 C1 subunits were found.

configurations, and the presented results provide a plausibility proof for the model of V4.

4.2 Results

4.2.1 Selectivity for Boundary Conformation

C2 units in the model can reproduce the selectivity of V4 neuronal responses. Model neurons reproduce the variety of selectivity described previously in V4 [Pasupathy and Connor, 2001], including selectivity to angular position and the curvature of boundary fragments. Fig. 4-3 compares the responses of an example V4 neuron to the corresponding C2 unit. This V4 neuron is selective for sharp convex boundary fragments positioned near the upper right corner of a stimulus, as shown in the response-magnitude ranked illustration of the stimuli in Fig. 4-3A. The modeled responses correspond closely to the physiological responses (coefficient of correlation $r = 0.91$; note that fitting V4 neural selectivity with a C2 unit is a more difficult problem than fitting selectivity at the S2 level, because the invariance operation, or pooling, of the C2 unit may cause interference between the selectivities of translated S2 units). This type of selectivity is achieved by the S2 configuration of 18 C1 subunits, shown schematically in Fig. 4-3C, which form a nonlinear template for the critical boundary fragments. The configuration of the C1 subunits offers a straightforward explanation for the observed selectivity. The C2 unit has a C1 subunit at 45° with a high weight, oriented along the radial direction (also at 45°) with respect to the center of the receptive field. This subunit configuration results in selectivity for sharp projections at 45° within the stimulus set and is described by the boundary conformation model as tuning for high curvature at 45° relative to the object center (more detailed comparison with the curvature and angular position tuning model will be given later).

C2 units can also reproduce selectivity for concave boundary fragments. Responses of the second example neuron, Fig. 4-4, exhibit selectivity for concave curvatures in the lower part of a stimulus. Again, there is a strong correspondence between the modeled and measured responses ($r = 0.91$). In this example, selectivity was achieved by a more complex S2 configuration with 21 oriented subunits, shown schematically in Fig. 4-4C. Note that there are two separated subunits with the strongest synaptic weights in the lower portion of the receptive field at -45° and 0° orientations; these correspond to boundary fragments found in many of the preferred stimuli. In general, the geometric configuration of oriented subunits in the model closely resembles the shape of a critical region in the stimuli that elicit high responses.

4.2.2 Population Analysis on the Selectivity for Boundary Conformation

Model C2 units can successfully fit the V4 population selectivity data and can generalize to V4 responses outside the training set. For each V4 neuron, all stimulus-response pairs were randomly splitted into two non-overlapping groups (a training and a testing set) in a standard cross-validation procedure. Fig. 4-5 shows correlation coefficient histograms for training and testing over the population of V4 neurons. The median correlation coefficient between the neural data and the C2 unit responses was 0.72 on the training set, and 0.57 on the test set over 6-fold cross-validation splits of the dataset. However, because the stimulus set is inevitably correlated, the test set correlation coefficients may be inflated.

Much of the variance in V4 neuron responses may be unexplainable due to noise or uncontrolled factors. Pasupathy and Connor (2001) estimated the noise variance by calculating

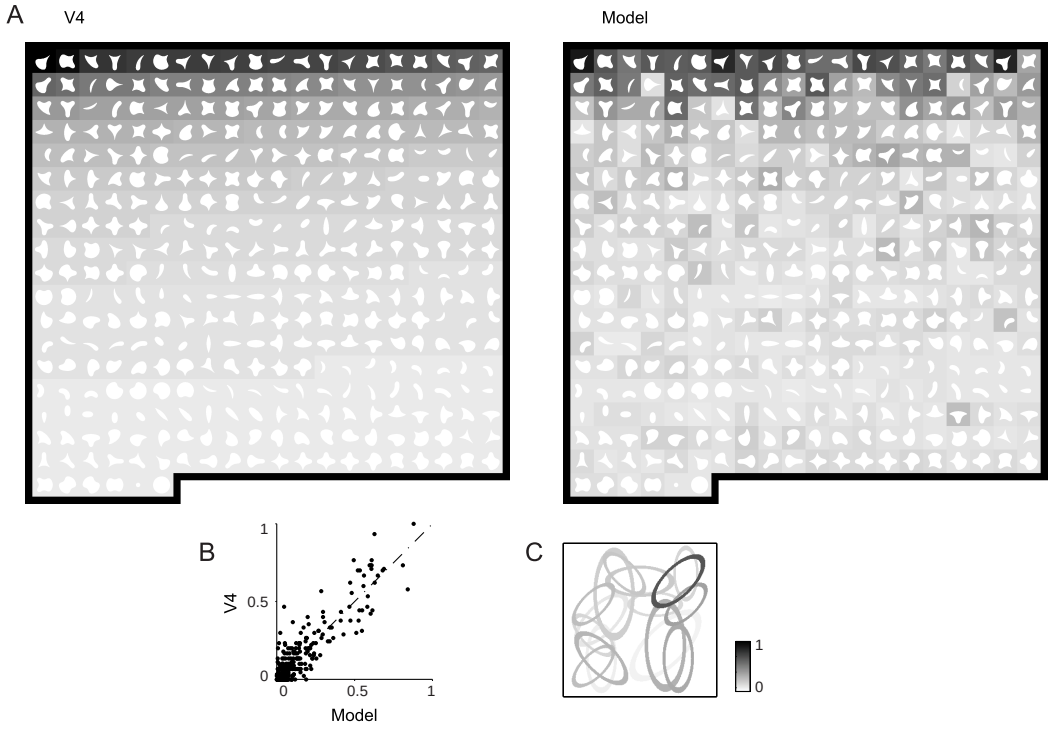


Figure 4-3: Comparison of model responses to a V4 neuron tuned to convex curvature. (A) On the left, the selectivity of a V4 neuron over 366 boundary conformation stimuli is shown, in order of decreasing response strength. The magnitude of the response is indicated by the gray scale (high response is darker). From the inspection of the response profile, it is apparent that this neuron is selective for a high convexity, or a sharp angle protruding out, on the upper right side of a stimulus. This is the same neuron that appears in Figure 5 of [Pasupathy and Connor, 2001]. The response of the C2 unit, modeling this V4 neuron's response, is shown on the right in the same stimulus order. A similar selectivity profile is observed. (B) The response of the V4 neuron is plotted against the model C2 unit's response for each stimulus. The goodness of fit, measured by the correlation coefficient, is 0.91 between this neuron and the model over the 366 boundary conformation stimuli. (C) The configuration of C1 subunits, projecting to S2 model units, is shown schematically. The configuration and weights of C1 afferents determine the selectivity of the S2 units and the resulting C2 unit. The locations and orientations of the C1 subunits are indicated by ellipses, and the strength of the synaptic weight is indicated by gray scale. This particular C2 unit is composed of S2 units each of which combines several C1 subunits with one strong afferent pointing diagonally outward in the upper right corner of the receptive field. This configuration is typical of C2 units that produce tuning to sharp curvature projections within the stimulus space.

the average expected squared differences across stimulus presentations. The estimated noise variance averaged 41.6% of the total variance. Using this estimate, on the training set the model accounted for 89% of the explainable variance ($r = 0.94$) and on the testing set the model accounted for 56% of the explainable variance ($r = 0.75$). Therefore, a large part of the explainable variance is described by the model. This result indicates that the model can generalize within the boundary conformation stimulus set.

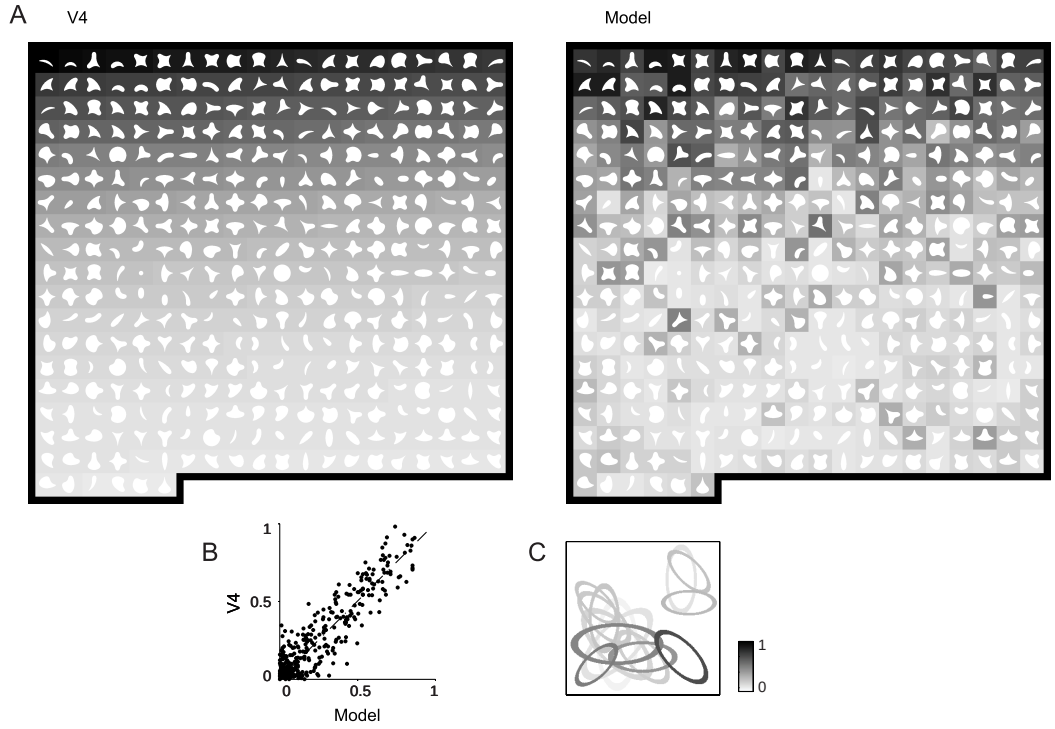


Figure 4-4: Comparison of model responses to a V4 neuron tuned to concave curvature. The selectivity of another example neuron is shown in the same format as Fig. 4-3. (A) This V4 neuron shows selectivity to boundary conformations with slightly concave curvature, or an obtuse angle, in the lower portion of the receptive field. (B) The model C2 unit closely matches the V4 neuron's response ($r = 0.91$). (C) The S2 configuration of the model is made up with 21 afferent C1 subunits. The two dominant subunits, oriented at -45° and 0° in the lower portion of the S2 receptive field, have a strong influence on the observed selectivity.

4.2.3 Invariance to Translation

The model not only matches V4 selectivity, but also produces V4 translation invariance. Responses of V4 neurons are invariant to translation (i.e., their selectivity is preserved over a local translation range), as reported in many studies [Desimone and Schein, 1987; Gallant et al., 1996; Pasupathy and Connor, 1999, 2001]. The population of C2 units used to fit the population of V4 neurons reproduced the selectivity of those V4 neurons, while still maintaining invariance to translation. Selectivity and invariance are two competing requirements and the model C2 units satisfy both requirements. The results in Fig. 4-6 show that the built-in invariance mechanism (at the level of C2) operates as expected, reproducing the observed translation invariance in the experimental data on the boundary conformation stimuli. Fig. 4-6A shows the invariance properties of the C2 unit from Fig. 4-3. Eight stimuli, which span the response range, are sampled across a 5x5 grid of positions with intervals equal to half the classical receptive field radius. Not only does the stimulus that produces a high response at the center of the receptive field produce high responses over a range of translation, but more importantly, the selectivity is preserved over translation (i.e., the ranking of the eight stimuli is preserved over translation within a given range). Fig. 4-6B and 4-6C show that the observed translation invariance of V4 neurons is captured

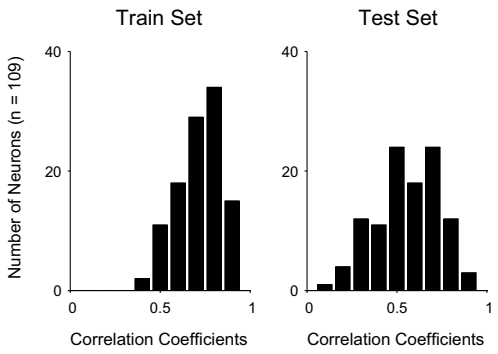


Figure 4-5: Generalization of the model to the stimuli outside of the training set. The model is able to predict the response of V4 neurons to boundary conformation stimuli not included in the training set. Using a 6-fold cross-validation methodology across the population of V4 neurons, a model C2 unit was determined for each V4 neuron using a training set, and the resulting model was used to predict the V4 neuron's response to the testing set. A histogram of correlation coefficients on the training (left) and testing (right) sets are shown. Over the population, the median correlation coefficients were 0.72 on the training set and 0.57 on the testing set. These numbers are based on the averages over the 6-fold cross-validation.

by the population of C2 units. Because the C2 units are selective for complex, nonlinear conjunctions of oriented features and the invariance operation is based on pooling from a discrete number of afferents, the translated stimuli sometimes result in changes of selectivity. A few C2 units in Fig. 4-6B show that translated non-optimal stimuli can produce greater responses; but on average, as shown in Fig. 4-6C, optimal stimuli within a range of translation produce stronger responses.

4.2.4 Responses to Bar and Grating Stimuli

The model is capable of reproducing the responses of individual V4 neurons to novel stimuli, such as bars and gratings. The population of C2 units produces responses that are consistent with the general findings that populations of V4 neurons show a wide range of orientation selectivity and bandwidths, that individual V4 neurons exhibit multiple peaks in their orientation tuning curves, and that V4 neurons show a strong preference for polar and hyperbolic gratings over Cartesian gratings.

To compute the orientation bandwidth of each C2 unit, the orientation selectivity of each model unit was measured using bar stimuli at various orientations (10° steps), widths (5, 10, 20, 30, and 50% of the receptive field size), and locations within the receptive field. The orientation bandwidth of each model C2 unit, the full width at half maximum response, with linear interpolation as in Figure 6A of [Desimone and Schein, 1987], was taken for the bar that produced the highest response across different locations. The multimodal nature of the orientation tuning curves was assessed using a bimodal tuning index, Eq. 9 in [David et al., 2006], after subtracting the C2 response to a blank stimulus.

Fig. 4-7A provides a summary plot of orientation bandwidths measured for 97 model C2 units (out of 109 C2 units, 97 had a response to a bar stimulus that was at least 10% of the maximum response to the contour stimulus set). The median orientation bandwidth for the C2 population was 52° (cf. around 74° in [Desimone and Schein, 1987; David et al., 2006]), and the distribution of the orientation bandwidths covers a wide range that is comparable

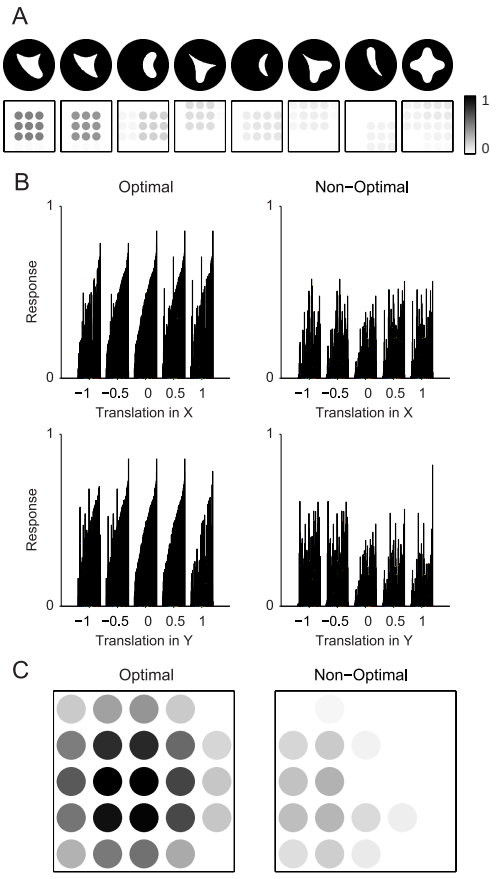


Figure 4-6: Translation invariance of model neurons. C2 units are invariant to translations, comparable to the invariance observed in V4. (A) On the top row, 8 different stimuli, which elicit a wide range of responses from the C2 unit of Fig. 4-3, are shown, and on the bottom row, the corresponding responses of this C2 unit are shown. Each stimulus was presented at 25 positions on a 5x5 grid (separated by half an S2 receptive field radius) centered on the C2 receptive field, following Figure 6A of [Pasupathy and Connor, 2001]. The center position in the 5x5 grid corresponds to the default presentation condition. Note that selectivity is preserved over a range of translation. (B) The population of 109 C2 units also shows invariance to translation and preserved selectivity over translation. The responses of the 109 C2 units to the stimulus that produced the highest response (optimal) and the stimulus that produced the lowest response (non-optimal) from the same set of 8 stimuli from (A) are displayed. Each thin bar corresponds to the response of a C2 unit, averaged along the orthogonal directions within the 5x5 grid, and the bars are sorted according to the responses at the default presentation condition. The x-axes are in units of S2 receptive field size. (C) This figure shows the average, normalized responses to the optimal and non-optimal stimuli, out of 8 shown in the top row of (A), across 109 C2 units for each stimulus position. The selectivity is generally preserved over translation. The responses to the optimal and non-optimal stimuli at the central position for each C2 unit are normalized to be 1 and 0 respectively, so that each unit makes equal contribution to this plot.

to the physiologically measured range.

Individual orientation tuning curves also indicated that many model C2 units were selective for multiple orientations. Such multi-modal orientation tuning, as opposed to the unimodal tuning in V1, is one of the characteristics of V4 neurons, and it arises naturally in this model because each model unit is composed of several differently oriented subunits. Although a number of model units have more than two peaks in their tuning curves, bimodal tuning indices, which characterize the deviation from the unimodal tuning behavior [David et al., 2006], were calculated. Fig. 4-7B presents a summary plot of bimodal tuning index for the 97 model C2 units that were responsive to the bar stimuli. The overall range of the bimodal index distribution in Fig. 4-7B is comparable with Figure 5D in David et al. (2006) and has a similar profile: a peak near zero with a dramatic falloff as the bimodal index increases. The median bimodal index over the population of C2 units was 0.12, comparable with 0.09 over the V4 population measured in David et al. (2006).

In order to test individual C2 units to grating stimuli, the same 109 model C2 units, fit to the V4 population, were presented with three types of gratings: 30 Cartesian, 40 polar, and 20 hyperbolic gratings each at 4 different phases to reproduce the stimulus set used in Gallant et al. (1996). The boundary conformation stimuli produced an average response of 0.22 from 109 C2 units, whereas the polar and hyperbolic grating stimuli produced an average response of 0.14 (1.0 is the maximum measured response over the main boundary conformation stimulus set). However, for 39% of the C2 units, the most preferred stimulus was one of the grating stimuli and not one of the boundary conformation stimuli. This result suggests that some V4 neurons selective for curved object boundary fragments might also show significantly higher responses to grating stimuli and other textures.

In correspondence with the report of a distinct group of V4 neurons that are highly selective for hyperbolic gratings [Gallant et al., 1996], some C2 units within this population are also highly selective for hyperbolic gratings. For example the C2 unit used to model the V4 neuron in Fig. 4-3 showed a strong preference for hyperbolic gratings, as its maximum response over hyperbolic gratings, 0.90, was much greater than the maximum responses over both polar gratings, 0.39, and Cartesian gratings, 0.04.

The population of C2 units also produces previously measured V4 population response characteristics to gratings. The distribution of grating class selectivity is shown in Fig. 4-7C. Quantitatively, mean responses to the preferred stimulus within each grating class were 0.0 for Cartesian, 0.16 for polar, and 0.20 for hyperbolic, qualitatively matching the finding of a population-wide bias toward non-Cartesian gratings [Gallant et al., 1996]. Many of the C2 units produced a maximal response to one grating class at least twice that of the other two classes: 1% for Cartesian, 35% for polar, and 26% for hyperbolic gratings. The reported experimental findings were 2%, 11%, and 10%, respectively.

The C2 population tends to be more strongly responsive to the non-Cartesian gratings than reported in [Gallant et al., 1996]. This discrepancy may be due to different screening processes used in the two experiments (V4 neurons in [Pasupathy and Connor, 2001] were recorded only if they responded to complex stimuli, and were skipped if they appeared responsive only to bar orientation). The C2 population also tends to show less selective responses between the polar and hyperbolic gratings than the neural data, as indicated by the concentrated points near the polar-hyperbolic grating boundary in Fig. 4-7C. An earlier modeling study [Kouh and Riesenhuber, 2003] suggests that a larger distance between the orientation-selective subunits can increase the variance of responses to these non-Cartesian grating classes, but this parameter was fixed in all simulations.

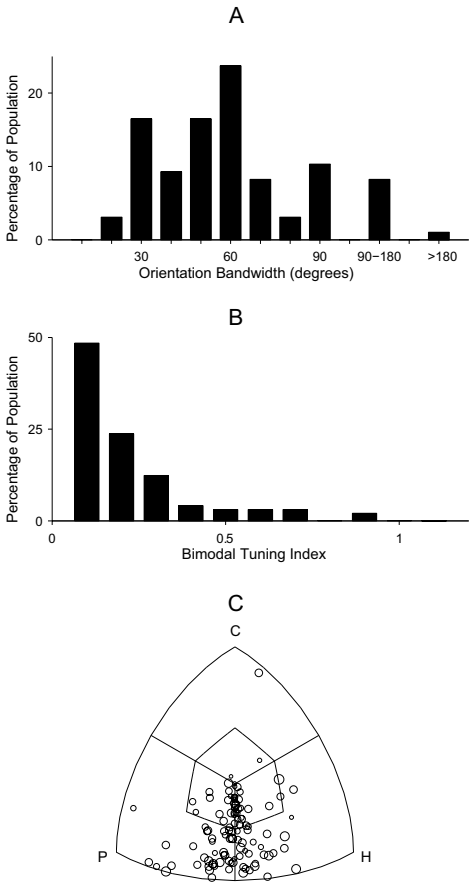


Figure 4-7: Testing selectivity of C2 units to bars and gratings. The responses from a population of C2 units, fit to the population of V4 neurons from [Pasupathy and Connor, 2001], are measured, using bars and gratings stimuli. (A) shows a histogram of orientation bandwidths measured for 97 model of the C2 units that showed significant response to bar stimuli. The median orientation bandwidth for the C2 population was 52° (cf. around 74° in [Desimone and Schein, 1987; David et al., 2006]), and the distribution of the orientation bandwidths covers a wide range that is comparable to the physiologically measured range. (B) shows a histogram over bimodal tuning index for the same 97 model C2 units. The overall range of the bimodal index distribution is comparable with Figure 5D in David et al. (2006) and has a similar profile. The median bimodal index over the population of C2 units was 0.12 (cf. 0.09 in David et al. (2006)). (C) shows a summary plot of the responses of all 109 C2 units to Cartesian, polar, and hyperbolic gratings. The stimuli, analysis procedure, and plotting convention follow the those of [Gallant et al., 1996]: For each C2 unit, the maximum responses to each grating class, as a 3-dimensional vector, were normalized to unit length and plotted in a 3-d space with each axis representing the response to a grating class (the viewpoint is oriented so that the origin of the coordinate system is at the center, and the vector whose responses are equal is pointing directly out of the page). The vector for each C2 unit is plotted as a circle, whose size indicates the magnitude of the maximum response over all grating stimuli. The results clear show a population bias toward the polar and hyperbolic gratings, which is an observed characteristic of area V4 [Gallant et al., 1996]. The differences in the degree of such bias may be due to different screening methods used in [Pasupathy and Connor, 2001] and [Gallant et al., 1996].

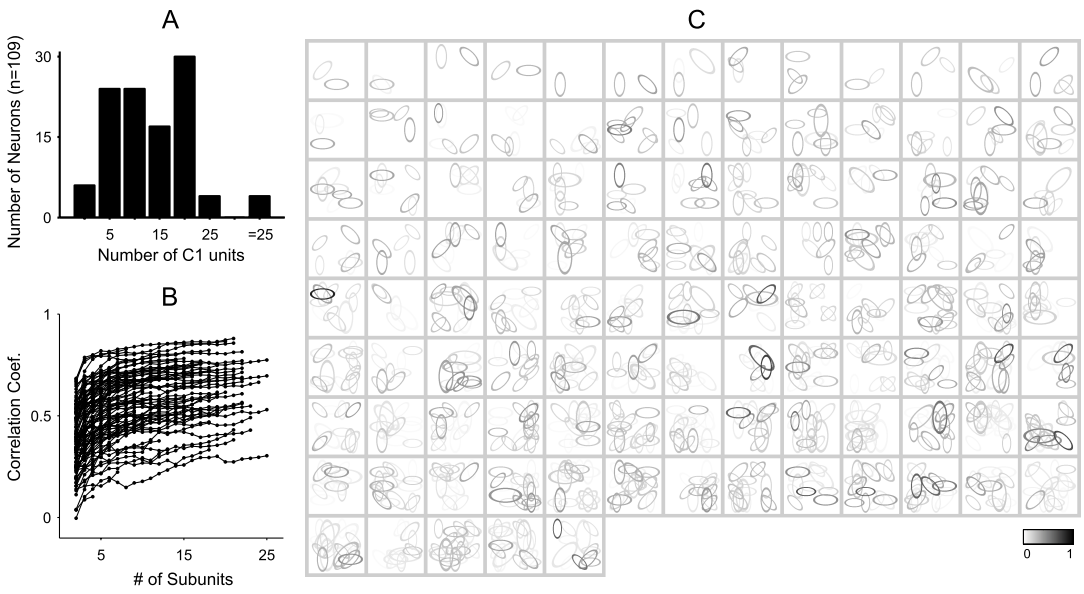


Figure 4-8: Complexity distribution of C2 units. (A) A varying number of afferent subunit is required to fit the 109 V4 neurons. The median was 13. The number of subunits was chosen to achieve the minimum average test error over 6-fold cross-validation. A maximum of 25 subunits was used in the fitting. (B) The evolution of the correlation coefficients (on the test set of the cross-validation) is shown along with the corresponding number of C1 afferent units. Individual V4 neuron, or the C2 unit modeling its response, is represented by each line, and based on the cross-validation criteria, fewer than 25 subunits are used for the final fit. (C) The S2 configurations for all 109 C2 units are shown in the order of increasing number of afferents. This figure illustrates that a population of V4 neurons is likely to come in widely different configurations and combinations of afferent inputs.

4.2.5 Complexity of V4 Neurons

Based on the model, it is possible to estimate the complexity of the V4 neuron population (note, however, that the recorded V4 population may be subject to some selection biases). Fig. 4-8A shows a distribution of the number of C1 afferent units found by the cross-validation analysis. For comparison, the neurons from Fig. 4-3 and 4-4 required 18 and 21 subunits, respectively. The results for predicting stimuli outside the training set, Fig. 4-5, are based on this distribution of C1 subunits. The median number of C1 afferent units found for the distribution was 13. In other words, a median of 16 parameters (13 plus 3 parameters in the sigmoid function, Eq. 2.9) were required to explain the measured V4 responses to the boundary conformation stimulus set. Fig. 4-8B shows the evolution of the correlation coefficients of the predicted responses for each V4 neuron. Overall, the correlation coefficient continues to improve with each additional C1 afferent units, indicating that the fitting methodology does not over-fit the neural responses and, within the framework of the model, that these additional C1 afferents are necessary for estimating the complexity of V4 neurons. The fitted model suggests that V4 neurons are not homogeneous in their complexity, but span a continuum in their selectivity to complex stimuli. This continuum is illustrated by the S2 configuration diagrams of all 109 neurons in Fig. 4-8C.

4.2.6 Comparison with the Curvature and Angular Position Tuning Model

The C2 units in the model provide a mechanistic explanation of V4 selectivity, while in Pasupathy and Connor (2001), tuning functions on curvature and angular position of the boundary fragments provide another description of the response profiles of the recorded V4 neurons. Therefore, it is worthwhile to examine the correspondence between the configurations of S2 afferents with the tuning functions for curvature and angular position derived in Pasupathy and Connor (2001). The C2 model fits are compared with the 4D curvature and angular position tuning functions described in [Pasupathy and Connor, 2001], in terms of the goodness of fit (correlation coefficient), the peak locations of angular position, and the degree of curvature.

Both C2 units and 4D curvature-angular position tuning functions capture much of the response variance of V4 neurons. The median correlation coefficients of two slightly different curvature-angular position tuning models were 0.46 and 0.57 respectively [Pasupathy and Connor, 2001]. There is a high correspondence between the correlation coefficients found for C2 units and the curvature-angular position tuning fits, as shown in Fig. 4-9A. This may not be surprising, as both models produce tuning functions in the space of contour segments that make up these stimuli.

In many cases, there is an intuitive relationship between the geometric configuration of a C2 unit's oriented C1 afferents and the tuning parameters in curvature and angular position space (i.e., Fig. 4-3C, concave curvature tuning, and Fig. 4-4C, convex curvature tuning, show such correspondence at specific angular positions). This relationship was quantitatively investigated by examining the parameters at the S2 level and comparing them to the peak locations of angular position and the degree of curvature found with the parameterized tuning functions. It is found that the angular position tuning is closely related to the weighted average of subunit locations, illustrated in Fig. 4-9B. Because the receptive fields of S2 units are large in comparison to C2 units (S2 RF radius = 0.75 x C2 RF radius), any spatial bias in the C1 inputs to S2 units will create a spatial bias at the C2 level. If this spatial bias is concentrated, the C2 unit will have a "hot spot" in angular position space.

To compare model parameters with curvature tuning, two main cases were considered, based on the criterion of whether there was one dominant subunit or many. If the second largest weight was smaller than 70 percent of the largest weight, only the strongest subunit was examined (Fig. 4-9C). Otherwise, the largest two subunits are considered (Fig. 4-9D). The curvature tuning comparisons are further divided into two cases, based on the criterion of whether the absolute value of tuned curvature was higher or lower than 0.7 (as defined by the curvature scale in Pasupathy and Connor (2001)). Since curvature is defined as a change in tangential angle over arc length, the joint distributions of the differences in subunit orientations (roughly corresponding to the change in tangential angle) and the differences in angular positions of two subunits (roughly proportional to the arc length) are computed. There were only four discrete orientations for the C1 units in the model, and the orientation differences were binned by 0, 90, and 45/135 degrees (the differences of 45° and 135° are ill-defined). The angular position differences were binned by small, medium, and large differences (indicated by S, M, and L in the label) in 60° steps.

Fig. 4-9C and 4-9D show that some curvature tuning can be characterized by simple geometric relationships between C1 afferents. When there is one dominant subunit, its orientation has a strong influence on whether the neuron is tuned for sharp or broad curvature fragments. If the subunit orientation and its angular position are parallel (for example

see Fig. 4-3C), the neuron generally produces high responses to sharp curvature fragments, which is evident from the bias towards 0° in the top row of Fig. 4-9C. If they are orthogonal, then the neuron is generally tuned for low curvature values, which is evident from the bias towards 90° in the bottom row of Fig. 4-9C. When multiple subunits have strong weights (like the example neuron in Fig. 4-4), the differences in their orientations and angular positions affect the curvature tuning, since curvature is determined by the rate of change in the tangent angle over the arc length. For the low curvature-tuned neurons, the two strongest subunits tend to have different orientations, and the angular position differences (proportional to the arc length) tend to be large (top row of Fig. 4-9D).

Note that this analysis also shows that the correspondence between these two models is not always straightforward. For example, some neurons that exhibit tuning to high curvature and are fit with C2 units with one dominant C1 unit, have subunit orientations that are perpendicular to the radial direction instead of parallel. A full description of a C2 unit's tuning properties requires the inclusion of all the C1 afferents, and the approximations used here may not capture the full situation. Nonetheless, the geometric arrangement of oriented V1-like afferents (C1 units) can explain the observed curvature and angular position tuning behavior in many V4 neurons.

4.3 Discussion

The presented results demonstrate that a quantitative model of the ventral stream, theoretically motivated and biologically plausible, reproduces visual shape selectivity and invariance properties of area V4 from the known properties of lower visual area V1. The model achieves V4-like representations through a nonlinear, translation-invariant combination of locally selective subunits, suggesting a computational mechanism within or culminating in area V4. Modeling the responses of 109 V4 neurons measured with a boundary conformation stimulus set, the simulated C2 units successfully reproduce V4 selectivity and invariance to local translation. Over the population of measured V4 neurons, the model produces an average correlation coefficient of 0.57 (uncorrected for explainable variance) on test sets of V4 responses to boundary conformation stimuli. The population of C2 units qualitatively generalizes to other experimental stimulus sets using bars and complex gratings.

C2 units may form an intermediate code for representing boundary conformations in natural images. Fig. 4-10 shows the responses of the two C2 units presented in Fig. 4-3 and 4-4 to two natural images. Based on the observed tuning properties of these neurons, it is not surprising to see that the first C2 unit responds strongly to the upper fins in the dolphin images, which contain sharp convex projections toward the upper right direction. The second C2 unit, which is selective for concave fragments in the lower portion of its receptive field, yields strong responses to several such boundary elements within the dolphin images. The graded responses of C2 unit populations may then form a representation of natural images that is particularly tuned to the conformations of various contours within an image. This code may be equivalent to the description provided by a previous study which demonstrated how a population code of V4 tuning functions could effectively represent contour stimuli [Pasupathy and Connor, 2002]. As seen in the two example images here, C2 responses can represent complex shapes or objects, even when curves and edges are difficult to define or segment and when the informative features are embedded within the boundary of an object (e.g., eyes, mouth, and nose within a face). Demonstrating this point, C2 units have been used as visual features to perform robust object recognition in natural images

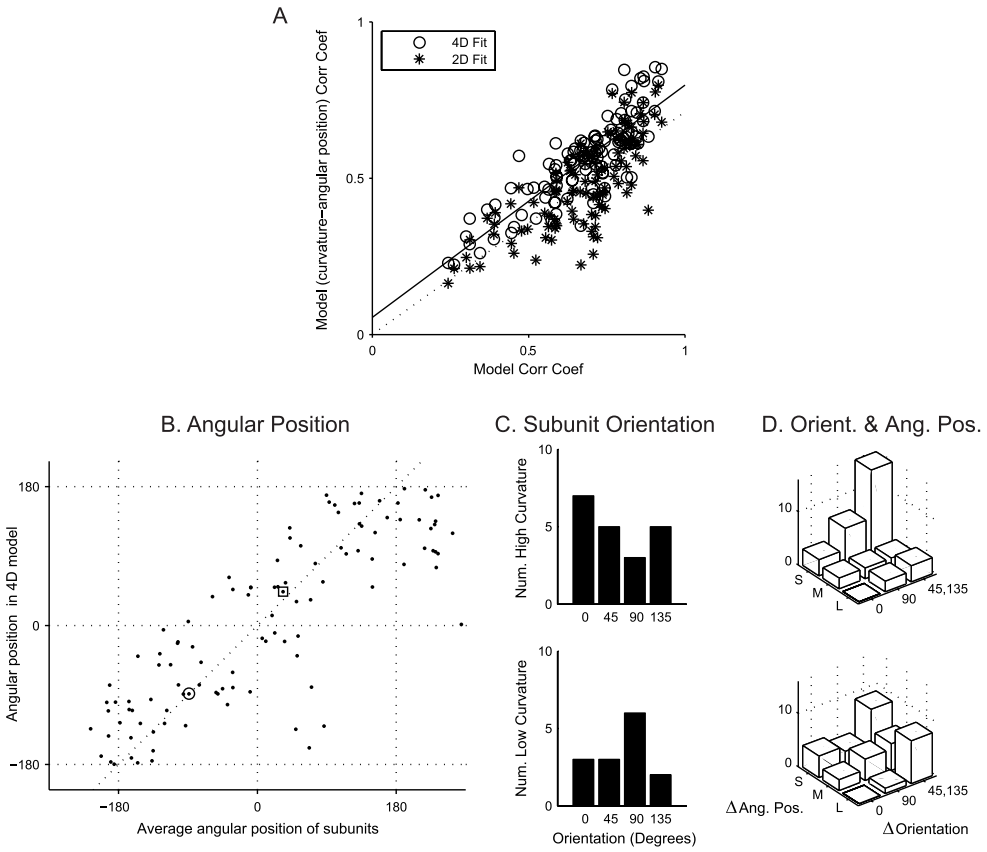


Figure 4-9: Comparison of the C2 model and the boundary conformation model. (A) Comparison of goodness of fits by the current V4 model (Fig. 4-1) and two boundary conformation tuning models (2D and 4D curvature and angular-position models) described in [Pasupathy and Connor, 2001]. (B) The angular position of the boundary conformation tuning function correlates with the "center of mass" of all subunits (weighted by synaptic weight). The example neurons of Fig. 4-3 and 4-4 are indicated by the square and circle symbols respectively (at 45° and -90°). (C) and (D) compare C1 subunit parameters with curvature tuning. Neurons are separated for this analysis into those that are tuned to high (top row) and low (bottom row) curvature values and models with one dominant subunit (first column) and many dominant subunits (second column). High curvature tuning can be achieved by a single dominant subunit oriented approximately radially, as seen in the first row of (C) (the subunit orientation with respect to its angular position is zero degree, similar to the example C2 unit in Fig. 4-3). If the subunit orientation was at 90° with respect to its angular position, the C2 unit tends to be tuned to the low curvature values, as shown in the bottom row of (C). When there are multiple dominant subunits, the two strongest subunits are considered for simplicity in (D). The joint distributions for the difference in subunit orientations and angular positions (binned into small, S, medium, M, and large, L, differences) are shown. Low curvature tuning tends to arise from a large angular position separation (arc length) between the subunits, as indicated by the skewed (toward larger angular position differences) joint histogram in the bottom row of (D) (an example is the C2 unit in Fig. 4-4). The results indicate that although some trends of correspondence between these two different models can be identified, the correspondence is not always straightforward.

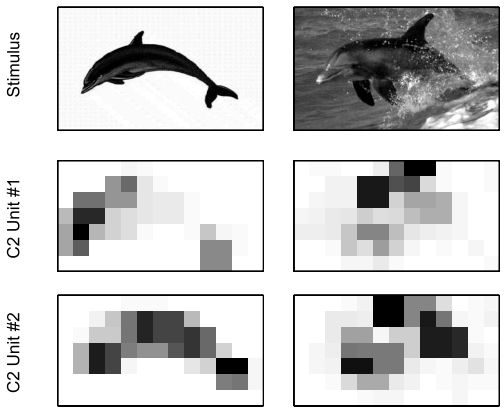


Figure 4-10: Model responses to natural images. Two images of dolphins (first row) and the responses of the two example C2 units from Fig. 4-3 (second row) and 4-4 (third row) to these images are shown. Because C2 receptive fields cover a small fraction of these images, the response of a C2 unit was calculated on overlapping (by a factor of 1/3 the C2 receptive field) crops of the stimulus image. Based on their shape selectivity, these model C2 units respond strongly to certain features within the image, as indicated by the gray-scale response maps in the second and third rows (dark areas indicate a high response). The images are from Fei-Fei et al. (2004).

[Serre et al., 2005, 2007a,b]. These results may suggest that V4 model neurons can respond like, and therefore be considered as, boundary conformation filters, just as V1 neurons can be considered edge or orientation filters [Chisum and Fitzpatrick, 2004; Daugman, 1980; Jones and Palmer, 1987; Mahon and De Valois, 2001; Ringach, 2004].

The current model of V4 is congruent with the major findings in Gallant et al. (1996), which indicate a bias within the population of V4 neurons to non-Cartesian gratings. Gallant et al. (1996) also proposed a mechanism, analogous to the simple to complex cell transformation in V1 proposed by Hubel and Wiesel (1962), to account for V4 responses. The ability of the model to predict the responses of a novel stimulus class given the responses of a training stimulus set, suggests a possible future test for the model: test the grating selectivity predictions of C2 units, which were derived from V4 measurements using boundary conformation stimuli, against the physiologically measured responses of these same V4 neurons to gratings. In addition, this model of V4 can be shown [Serre et al., 2005] to reproduce the experimental data of Reynolds et al. (1999) for the condition without attentional modulation in V4. While the model is not designed to address some other known properties of V4 responses, namely spectral and color selectivity [Schein and Desimone, 1990], 3-dimensional orientation tuning [Hinkle and Connor, 2002], saliency [Mazer and Gallant, 2003], or attentional effects [Reynolds et al., 1999], it accounts for much of the structural object dependent selectivity and invariance currently described.

A recent publication [David et al., 2006] proposed that V4 response properties could be described with a second-order nonlinearity, called the spectral receptive field (SRF). This description of V4 neurons is phenomenological and aimed at providing a robust regression model of the neural response, while the current model is motivated and constrained by the computational goal of explaining object recognition in the ventral stream. It is therefore interesting to ask whether a connection exists between the two descriptions at the level of V4 cells. In fact, Volterra series analysis reveals that the leading term of the current

model is similar to the SRF (involving the spectral power of the input pattern), but the series associated with this model contains additional terms that are not negligible [Cadieu et al., 2007]. In this sense, the model described here could be considered similar but not identical to the model of David et al. The additional aspects of the model describe some important aspects of V4 responses that are not described by the SRF. Because the SRF model lacks the spatial organization of afferent inputs, its response profiles will not be (1) selective for angular position tuning, (2) sensitive to the relative positions of features in space, or (3) inhomogeneous within the receptive field, which are all attributes of C2 units. However, while the C2 model assumes a specific type of architecture and a set of nonlinear operations to explain the properties of the V4 neurons, the SRF model provides a more general and agnostic regression framework, which can be used to analyze and predict the neural responses, not just specific to V4. The two models should ultimately be evaluated against experimental data. The correlation between predicted and actual data for the two models (0.32 for David et al. (2006) and 0.57 for this model) cannot be directly compared because the stimulus set used in David et al. (2006) is more complex and varied.

Learning may also play a critical role in the selectivity of V4 neurons. In the full model of the ventral pathway (see right side of Fig. 4-1) the configurations and weights between S2 units and their oriented C1 afferents, which determine the selectivity of the C2 units, are learned from natural viewing experiences by a simple, unsupervised learning mechanism. According to simulations, such learning mechanisms are capable of generating rich intermediate feature selectivities that account for the observed selectivity of V4 neurons [Serre et al., 2005, 2007a]. Building upon such intermediate feature selectivity, the model of the ventral pathway can perform object recognition tasks on natural images at performance levels at least as good as state-of-the-art image recognition algorithms and can mimic human performance in rapid categorization tasks [Serre et al., 2005, 2007a,b]. The invariance may also be learned in a biophysically plausible way (e.g., [Földiák, 1991; Wallis, 1996; Wiskott and Sejnowski, 2002]), during a developmental period, from natural viewing experiences, such as watching a temporal sequence of moving objects. If temporally correlated neurons in a neighborhood connect to the same higher order cell, the appropriate connectivity found between S2 and C2 units in the model can be generated [Serre et al., 2005].

How do V2 and a subpopulation of V1 neurons with more complex shape selectivity (beyond simple orientation tuning) fit into this model of V4? Several studies [Mahon and De Valois, 2001; Hegde and Van Essen, 2007] have found that rather complex shape selectivities are already present in V1, but there are relatively few experimental and theoretical studies of V2 and such hyper-complex V1 neurons, making it difficult to include concrete constraints in the analysis. However, three hypotheses about their roles and functions are suggested by the current hierarchical model. (1) The selectivity and invariance seen in V4 may be constructed from yet another intermediate representation, which itself is both more selective and more invariant than simple orientation-selectivity, but less selective and less invariant than V4 (producing a continuum of receptive field sizes and invariance ranges depending on pooling ranges within the model); or (2) some hyper-complex V1 and V2 neurons are analogous to S2 units of the model, so that they have complex shape selectivity, but weak translation invariance. The more invariant representation is realized by a V4 neuron pooling over these V1 or V2 neurons. Under this hypothesis, if S2-like representation is highly concentrated in V2, the cortico-cortical projections between areas V2 and V4 would represent fundamentally different transformations from the projections between V1 and V2; or (3) Area V2 is representationally similar to V1 for feedforward responses. Under this last hypothesis, area V4 may contain neurons analogous to both S2 and C2 units in the model,

or the selectivity representations (of S2 units) are computed through dendritic computations within neurons in V4 [Mel et al., 1998; Zhang et al., 1993]. Experimental findings show that the majority of measured V4 responses are invariant to local translation, supporting the hypotheses that S2-like selectivity representations with small invariance range are present in another area of the brain, that they are computed implicitly in V4, or that there has been an experimental sampling bias. However, although V2 neurons are known to show selectivity over a range of stimulus sets [Hegde and Van Essen, 2003], there is not enough experimental data so far to verify or even distinguish these hypotheses. Carefully measuring and comparing both selectivity and invariance of areas V2 and V4 would be necessary to resolve this issue.

The V4 dataset from Pasupathy and Connor (2001) contained recordings using only one stimulus class and did not allow us to test the generalization abilities of the model to other types of stimuli. Although attempts were made to gauge the generalization capacity of the model (using cross-validation within the boundary conformation stimulus set and observing model responses to gratings and natural images), the ultimate test will require a thorough examination across a wider range of stimulus sets, including natural images. Furthermore, the current model is applicable only to the rapid response of V4 neurons and does not explain attentional or top-down factors [Mazer and Gallant, 2003; Reynolds et al., 1999].

The analysis of the representations in V4, adds to the mounting evidence for canonical circuits present within the visual system. Interestingly, the proposed mechanism for selectivity in V4 (a normalized weighted summation over the inputs, Eq. 2.3) is quite similar to the model of MT cells proposed in a recent publication [Rust et al., 2006]. In addition, another recent study claims that motion integration in MT requires a local mechanism [Majaj et al., 2007], which may be analogous to the locally selective S2 units and more "global" C2 units for describing V4. Consequently, the same tuning and invariance operations may also be operating along the dorsal stream and may have a key role in determining various properties of motion selective neurons in MT. This model of V4 is also consistent with widely held beliefs on the ventral pathway, where more complex selectivity and a greater range of invariance properties are thought to be generated by precise combinations of afferent inputs. Previous quantitative studies have argued for similar mechanisms in other parts of the ventral stream [Perrett and Oram, 1993]. Further experimental work using parameterized shape spaces has shown that IT responses can be explained as a combination of invariant V4-like representations [Brincat and Connor, 2004], which is consistent with this model [Serre et al., 2005]. It has also been suggested that a tuning operation, used repeatedly in the current model, may be a suitable mechanism for producing generalization, a key attribute of any learning system [Poggio and Bizzi, 2004]. Therefore, instead of a collection of unrelated areas performing distinct tasks, the ventral pathway may be a system organized around two basic computational mechanisms necessary for robust object recognition.

Chapter 5

Comparisons with the IT Neurons

The inferior temporal (IT) cortex is the last purely visual area in the hierarchy of the primate visual cortex and contains neurons that are highly responsive to complex stimuli like images of a face or a hand [Gross et al., 1972]. Further electrophysiological studies have confirmed and extended the view of IT as the locus of the highly selective and invariant neurons that play an important role in the visual object recognition process.

In this chapter, the responses of the model units in the top layer, corresponding to the inferior temporal cortex, are examined, showing (1) the selective and invariant neural response properties, (2) the robust object recognition performance, and (3) the trade-off behavior between selectivity and invariance. Some of these results have been reported earlier with the idealized operations (i.e., the exact Gaussian and maximum operations) in [Riesenhuber and Poggio, 1999b; Hung et al., 2005], but the results shown here are obtained with the biologically plausible versions of these operations, corresponding to Eq. 2.3 and the neural circuits A or B in Fig. 2-1.[1]

5.1 Selective and Invariant Neural Responses

In the study of [Logothetis et al., 1995], monkeys were extensively trained to recognize a set of novel 3-dimensional 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 different 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 of (1) novel object class that the monkeys had not had any visual experience with and (2) the distractor objects, which allow the definition of the range of selective invariance. Simply measuring how much the neural response changes to the object transformation is not as meaningful

[1]For most simulations, the parameters in Eq. 2.3 were fixed at $(p, q, r) = (1, 2, 1)$ for the Gaussian-like and $(3, 2, 1)$ for the max-like operations all throughout the hierarchy, unless specified otherwise. Sigmoid functions have been applied after each Gaussian-like operation to span a reasonable operating range for the neural responses. Within each layer, the sigmoid parameters are fixed across different neural units. The responses of the top layer in the model are obtained by the cascade of the Gaussian-, max-, Gaussian-, max-, and Gaussian-like operations on the stimuli (5 layers, corresponding to the original model [Riesenhuber and Poggio, 1999b] and "bypass" route in [Serre et al., 2005]).

as making a comparison to a reference value that is defined by the maximum response to many distractor objects. Such comparison provides a measure of the range of invariance to the transformed versions of the preferred object.[2]

The model demonstrates quantitatively that the combination of the Gaussian-like and max-like operations can achieve view-tuned, selective and invariant tuning properties. Fig. 5-1 shows the responses of one model unit from the top layer, which was created by making it selective for a particular view of a paperclip stimulus. As the object is slowly transformed away from the preferred view, by 3-dimensional rotation, scale or translation, the response of this model unit decreases. However, over some ranges of these transformations, its response is greater than the responses to other distractor paperclips. Therefore, this model unit achieves the same selective and invariant response properties as in [Riesenhuber and Poggio, 1999b; Serre et al., 2005], by using just one class of transfer functions computed from realistic neural circuits. It can encode the presence of a particular paperclip, within these ranges of rotation, scaling and translation.

5.2 Object Recognition Performance

With their selective and invariant response properties, the inferior temporal neurons are believed to play a critical role in the visual object recognition process. In a recent study [Hung et al., 2005], it has been shown that a population of the IT neurons carries rich information about the object identities and categories. More specifically, by entering their responses (e.g., spike counts, local field potentials, etc.) into a classifier, it is possible to read out the identity and the category of a given object.

Fig. 5-2 shows the result of performing the same readout experiment as in [Hung et al., 2005], using the same set of stimuli and the same protocols for training and testing a support vector machine (SVM) classifier on the responses of the model IT neurons. The performance is significantly above the chance level, and closely resembles the performance obtained with the neural data. The readout performance (for both model and neural data) is invariant to the changes in size and position of the stimuli. In other words, the classifier can be trained on the responses to the stimuli at one particular position and size, and it will still perform well even when tested on the responses to the stimuli presented at some different positions and sizes. When the classifier is trained on the response of the model units from the lower layers in the hierarchy, the recognition performance is significantly less.

The results from the readout experiment, in conjunction with the invariant and selective response properties of the model units, suggest that the feedforward, hierarchical combi-

[2]Let y_A be the neural response to a stimulus A. If the neuron is selective for A, y_A will be different from y_B, the response to a different stimulus B. If the neuron is invariant to $T(A)$, a transformed or appearance-altered version of the stimulus A, its response will be the same (i.e., $y_A = y_{T(A)}$). However, in reality, it would be sufficient to have "similar enough" responses, so that the change in the output due to the transformation is smaller than the difference in the outputs to the target and other non-target stimuli. The range of selectivity and invariance, then, can be defined by the degree of stimulus transformation that satisfies the condition, $|y_A - y_{T(A)}| < |y_A - y_B|$. Beyond that point, the outputs are no longer selective for the stimulus A nor invariant (or tolerant) to the transformation T, against the non-target stimulus B. Based on the response of this single neuron, the object A can no longer be distinguished against another object B, under the appearance-altering transformation T. However, the identity of an object is not at all likely to be encoded by a single neuron (like a "grandmother cell"), but by a population. Therefore, even when the response of a single neuron is strongly perturbed by some object transformations (or by the cluttering objects within the receptive field), object recognition can be performed with the responses from multiple neurons (cf. readout experiment of [Hung et al., 2005]).

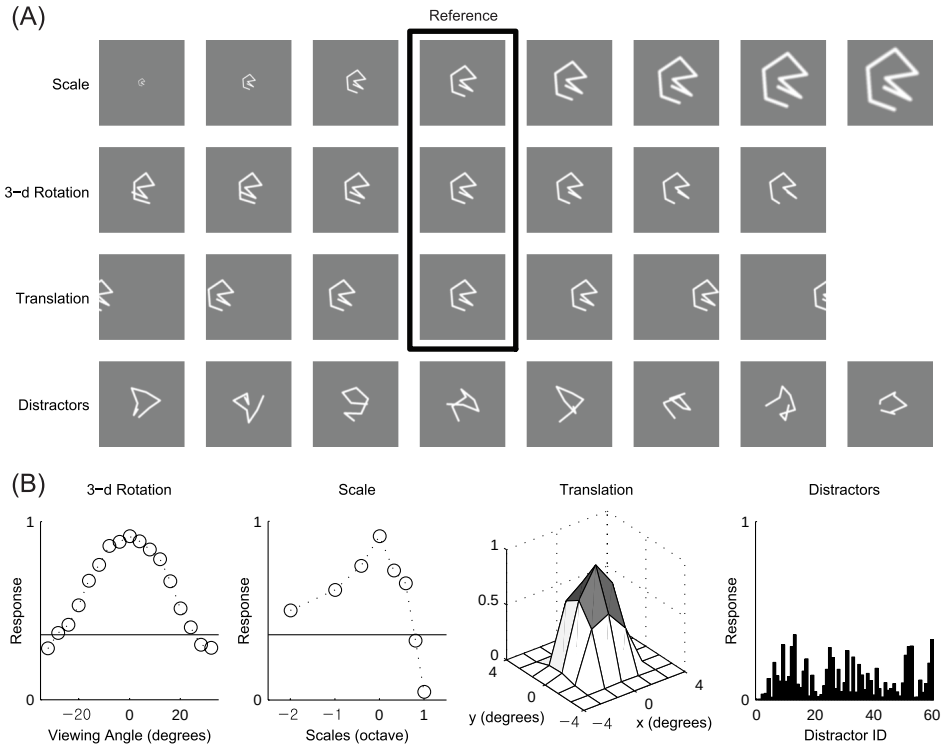


Figure 5-1: (A) Samples of the paperclip stimuli are shown. The reference paperclip (the images in the 4th column of the first 3 rows) is scaled, rotated in depth, and translated. The fourth row shows some examples of the distractor paperclips. (B) Tuning curves of a model inferior temporal neuron under 3D rotation, scale and translation of the preferred paperclip object, and the responses to 60 distractor stimuli are shown, as in [Logothetis et al., 1995; Riesenhuber and Poggio, 1999b]. The maximum distractor response is indicated by the horizontal lines in the first two panels. For the rotation experiment, the viewpoints were varied in the 4°steps. For scale, the stimulus size varied from 32 to 256 pixels in 32-pixel steps (32 pixels roughly correspond to one degree of the visual angle). For translation, the stimulus was moved across 8 degrees of the visual angle. In this simulation, the model units in the intermediate layer have been learned from an independent set of the paperclip images, because the responses of the model units learned from the natural images tend to be too small, resulting in very small invariance ranges. This model unit shows the rotational invariance range of around 60 degrees, scale invariance range over 2 octaves, and translation invariance range of 4 degrees of the visual angle. Other model units tuned to other reference paperclips show the similar results. Average invariance ranges for 20 random paperclip-tuned units were 54±8 degrees, 1.4±0.2 octaves, 3.8±0.5 and 4.0±0.4 degrees for rotation, scale and translation (x and y directions), in a good agreement with the experimentally observed values. (The model can produce somewhat larger scale invariance ranges, 2.3±0.3 octaves, when the exact maximum operation is used.)

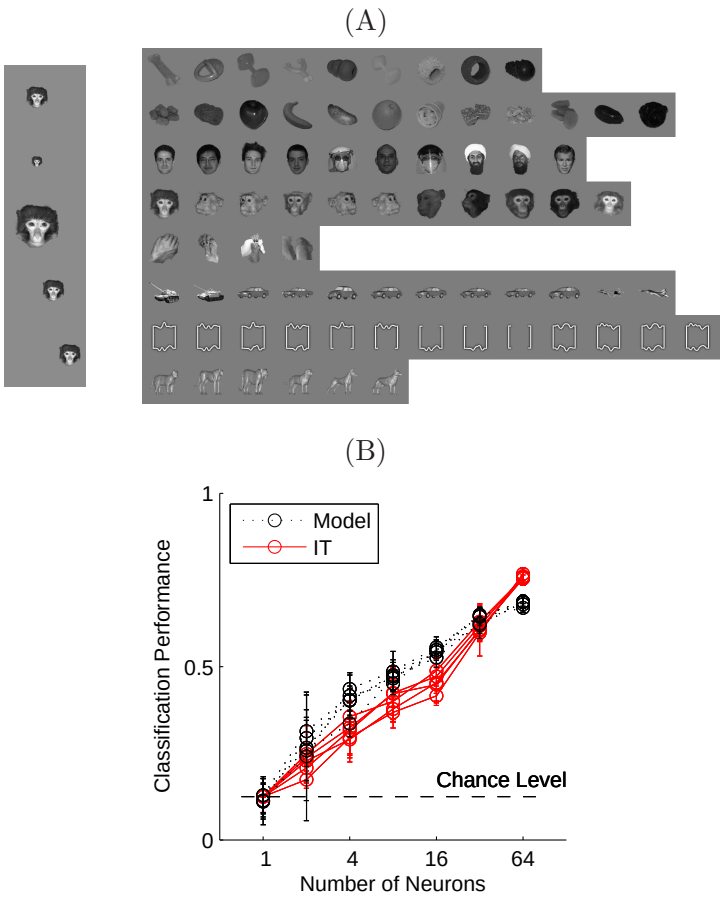


Figure 5-2: (A) The stimulus set for the read-out experiment is composed of 78 objects (right panel) in 8 categories (toy, food, human face, monkey face, hand, vehicle, box, and animal), presented in 5 different conditions (left panel): (1) default position and size, (2) half the size, (3) twice the size, (4) 2 degrees and (5) 4 degrees of the visual angle translation, exactly as in [Hung et al., 2005]. (B) A support vector machine classifier was trained on the responses to the stimulus images at one particular condition and tested on the responses to the other four conditions, resulting in 5 different performance curves for each dataset based on either the neural or the model responses. The performance of the classifier at varying number of the inputs is plotted. Note that the performance is clearly above the chance level and the 5 performance curves overlap, indicating that the neural and model responses are category-specific and invariant to the scale and translation transformations. Each set of training and testing was performed by randomly selecting a neural population from a pool of either 70 recorded inferior temporal neurons (solid lines) or 2000 model units (dashed lines). The error bars indicate the standard deviations on five independent random selections. Each model unit is tuned to a randomly chosen input pattern and receives 100 inputs from afferent units in the lower layer (analogous to the area V4 or PIT), which are tuned to and learned from the patches of the natural images. The performance by the model responses is quite close to the performance by the neural data, but seems to saturate, indicating that the pool of the 2000 model units (or their afferent units) is probably not diverse enough (i.e., an additional model unit carries less and redundant category information than the recorded inferior temporal neurons). The stimuli, neural data, and classification codes are provided by C. Hung, J. DiCarlo, and G. Kreiman [Hung et al., 2005]. Also see [Serre et al., 2005].

nation of the Gaussian-like and max-like operations computed by the biologically plausible neural circuit is sufficient for the robust object recognition. Other simulations have shown that the model can also perform quite well in more complex object recognition task, such as detecting the presence of an animal in the natural images, at a level comparable to humans (for short exposure times) and other state-of-the-art machine vision systems [Serre et al., 2007b].

5.3 Clutter Effects and Tradeoff between Selectivity and Invariance

One of the hallmarks of the primate visual system is its ability to rapidly and robustly detect and recognize an object under many challenging conditions. In particular, the visual system is capable of recognizing an object even when other objects are simultaneously present in the visual scene (e.g., finding a key on a cluttered desktop). From many electrophysiological studies (for example, [Reynolds et al., 1999; Gawne and Martin, 2002; Zoccolan et al., 2005, 2007]), it has been shown that clutter conditions can produce various modulatory effects on the neural responses in different visual areas.

In this section, the model of the ventral pathway [Serre et al., 2005] was used as an exploratory tool to account for the wide range of clutter effects with a single, unifying framework and to understand possible neural mechanisms and factors underlying these effects. Because the model deals with the feedforward neural responses (in the absence of attention or back-projecting signals), the application of the model is appropriate only for those experiments where the monkeys were performing a fixation task only (or an orthogonal, unrelated task as in [Zoccolan et al., 2007]) during a rapid presentation of the stimuli.

The work described here is based on close collaborations with D. Zoccolan and J. DiCarlo, who have performed the physiological experiments, and more details about the experiment can be found in [Zoccolan et al., 2007].

5.3.1 Experimental Findings

Object recognition and understanding of a scene gist can be accomplished in a fraction of a second even when multiple objects are simultaneously present in a scene [Thorpe et al., 1996; Hung et al., 2005; Li et al., 2002; Serre et al., 2007a], although the mental resource to process a cluttered scene is certainly not unlimited.

Many electrophysiological studies in different cortical areas have revealed that the neural response to multiple objects can be quite different from the responses to single, isolated objects. An introduction of a cluttering stimulus typically brings the response to an intermediate value between the responses to the stimuli in isolation [Zoccolan et al., 2005, 2007; Reynolds et al., 1999]. Quantitatively, such "clutter" or interference effects on the neural responses are quite variable: Some neurons show highly clutter-tolerant behaviors [Lampl et al., 2004; Gawne and Martin, 2002], average-like behaviors [Zoccolan et al., 2005; Reynolds et al., 1999], and a spectrum of other responses that may also depend on the stimuli (e.g., location between the stimuli as in [Livingstone and Conway, 2003; Freiwald et al., 2004]). The attentional state of the subject can also modulate the clutter response [Reynolds et al., 1999].

In a recent systematic study of selectivity and clutter tolerance of the IT neurons [Zoccolan et al., 2007], it was again found that there is a wide variation of selectivity and tolerance properties in the neural population. Some neurons were highly selective, generating a strong response to a specific object and low responses to other stimuli. Other neurons were less selective and generated sizable responses to many different stimuli. Hence, there was a distribution in the sparseness of the response profiles within the measured IT population. When two objects were presented together, the neurons typically produced an intermediate level of responses, compared to the responses to the isolated single objects. When an effective and an ineffective stimuli were paired together, typically the resulting response was lower than the response to the effective stimulus alone, but still higher than the response to the ineffective stimulus alone. Within the population, the effect of the clutter or the level of clutter tolerance varied widely. Some neurons maintained a high response even when an ineffective stimulus was paired with an effective stimulus, while other neurons showed large changes in the response levels. These results are consistent with other reports on the clutter effects [Zoccolan et al., 2005; Reynolds et al., 1999].

Interestingly, this study [Zoccolan et al., 2007] has also found that the neurons with high selectivity typically have low tolerance, and vice versa. In other words, the selectivity and tolerance properties of a neuron are correlated, so that high selectivity typically implies low tolerance to the clutter, and high tolerance implies low selectivity. Such a tradeoff between selectivity and tolerance was found to be very robust and independent of how selectivity and tolerance were measured. Furthermore, similar tradeoff behaviors were observed not just for the clutter, but for other types of stimulus transformations like position, size, and contrast changes.

5.3.2 Simulation with the Model

In the simulations, the experimental protocols of [Zoccolan et al., 2007] are closely followed, by using the same stimuli and the metrics for characterizing the sparseness and tolerance properties of neural response. Fig. 5-3 shows the stimulus set used in the experiment.

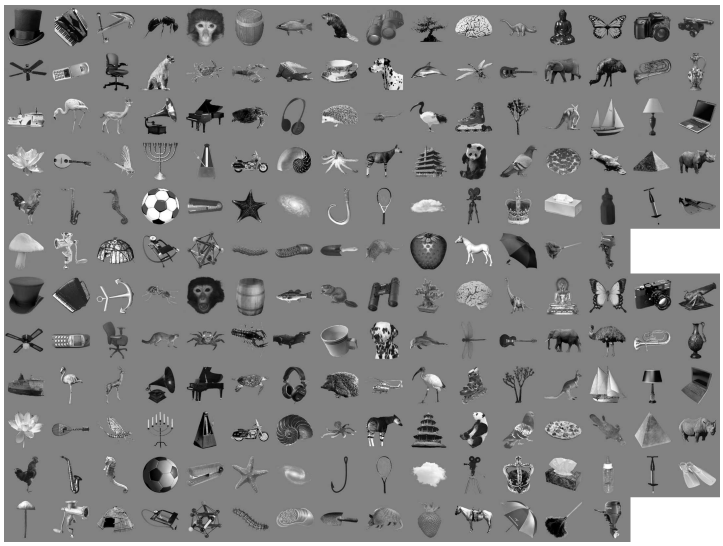
The selectivity of a neuron is based on its response across different stimulus identities (i.e., what is the neural response to stimulus X vs. stimulus Y?) [Vinje and Gallant, 2000]. High selectivity indicates that the neuron is sensitive to the identity of the stimulus, or it has a sharp tuning curve, or its response is sparse. The clutter tolerance or invariance of a neuron is based on the modulation of its responses to clutter (i.e., how does the response change when stimulus X is presented together with stimulus Y, compared with the cases when stimulus X and Y are presented in isolation?). High clutter tolerance indicates that the neuron is insensitive or tolerant to the cluttering object.

$$\text{Sparseness Index} = \frac{1 - (\sum_i R(i)/n)^2/(\sum_i R(i)^2/n)}{1 - 1/n}, \tag{5.1}$$

$$\text{Clutter Tolerance Index} = \frac{1}{6}\sum_{j=1}^{6}\frac{R(i+j) - R(j)}{R(i) - R(j)}. \tag{5.2}$$

$R(i)$ denotes the response of a neuron or a model IT unit to the i-th stimulus. The total number of the stimuli is denoted by n. The sparseness index of 0 indicates that the responses to all n stimuli were identical (i.e., there was no selectivity), and 1 indicates that only one stimulus produced a large response and other stimuli produced zero or negligible response.

(A)



(B)

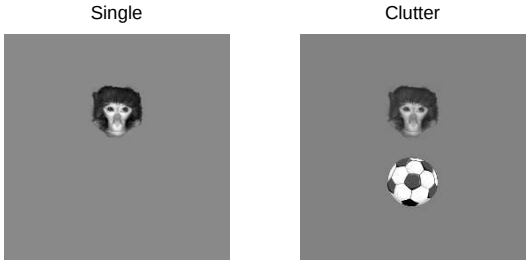


Figure 5-3: (A) Main stimulus set is composed of approximately 200 stimuli. There are two exemplars of each object. These objects are used to create single and clutter conditions, as shown in (B). The stimuli were provided by D. Zoccolan and J. DiCarlo.

Effectively, this index measures the ratio between the mean and the variance of the response profile over a set of stimuli, so that high variance is associated with high sparseness or selectivity, and low variance, with low selectivity. The response to the clutter is denoted by $R(i+j)$, and typically, the most effective stimulus i was paired with an ineffective stimulus j. Six such pairs are taken to compute an average clutter tolerance index of a neuron. The tolerance index of 1 indicates that the response was insensitive to clutter, and 0 indicates an extremely high sensitivity.

The S4 units in the model, corresponding to the AIT neurons studied in [Zoccolan et al., 2007], make synapses with the C2b units in the previous layer (see Fig. 3-1). The tuning properties of the S4 units are determined by the activation patterns of the randomly chosen afferent units in response to randomly chosen target objects (typically from the stimulus set). This means that a model IT neuron had, as a preferred stimulus, one of objects of the stimulus set. The presented simulation results are also similar when the preferred stimuli of the model IT units do not exactly match the objects from a fixed stimulus set (for example, when a small noise is added to the center of the tuning function).

The selectivity of the afferent units, corresponding to the neurons in V4 or PIT, is also learned from the stimulus set. Such an arrangement of using the features learned from the stimulus set tends to produce robust afferent responses and agree with the neural data better. In many simulations, the number of afferent input was varied, by choosing either random or top most activated afferents units (to the preferred stimulus of the efferent unit). The latter method of biasing the stronger afferents, as previously proposed and applied in other model simulations [Riesenhuber and Poggio, 1999b], can increase the tolerance to the clutter. This choice was based on the assumption that a neuron in inferior temporal cortex may preferentially strengthen its connections with those afferents that are more strongly activated during presentation of its optimal visual pattern (Hebb's rule). All other parameters in the lower layers of the model have been fixed throughout the simulations.

The specific parameters for the Gaussian-like operation used in the simulations were $(p, q, r) = (1, 2, 1)$ and $(\alpha, \beta) = (40, 0.9)$. Note that β determines the sensitivity or threshold of the output unit along each input dimension (i.e., it plays the similar role as σ in the Gaussian function). When β is large, the tuning function is narrow (small σ), and small deviations from the optimal input pattern (the center of the Gaussian-like function) result in a drastic reduction in the response of the output unit. On the other hand, when β is small, the tuning over the input space is broad (large σ) and the output unit can tolerate large deviations from the preferred pattern and still respond. For the max-like operation, $(p, q, r) = (3, 2, 1)$ and $k = 0$. Note that the response of a model unit is a value between 0 and 1, where 1 represents the maximum response of a neuron (e.g., maximal firing rate or spike count).

In addition to the clutter experiment, the response of the model IT units to other stimulus transformations was measured, by changing the position, size, and contrast of the presented stimuli, as in [Zoccolan et al., 2007]. It should be noted, however, that the ranges of position and size changes tested in the model do not exactly match those probed during the physiological experiments, because of the limited span of visual field (8 degrees) simulated in the model. Moreover, a more realistic and quantitative comparison with the data would require modeling the variations in the receptive field sizes and locations of inferior temporal neurons and the drop of retinal sampling as a function of retinal eccentricity. For the simulations on the contrast sensitivity of the model units, it was assumed that stimuli with lower contrast would produce smaller responses in the model units representing V1 neurons. Specifically, the responses of each S1 unit to the stimuli at 1.5, 2 and 3% con-

trast were reduced to 70, 80 and 90% of the response to the same stimulus at its reference contrast. The tolerance properties of each model unit were also assessed with the same indices as in [Zoccolan et al., 2007], by measuring the relative response reduction to these transformations.

As shown in Fig. 5-4, there are two main results, which are both consistent with the physiological data [Zoccolan et al., 2007].

1. There is a very wide range of, mostly suppressive, clutter effects, as shown by the wide distribution of the points along the ordinate in the bottom figures.

2. The selectivity (sparseness) and tolerance properties of the neural responses across the population show an inverse relationship.

In the following section, these two main effects are discussed in more detail.

5.3.3 Possible Mechanism of the Clutter Effects

In the model, the response of a neuron is determined by the pattern of inputs from the afferent neurons and the local neural circuitry. Depending on the type of the neural circuit and its parameters, the neural response may be Gaussian-like or max-like. The activation pattern of the afferent neurons is, in turn, generated by the responses of their own afferent neurons and their own local neural circuitries. In other words, a neural response is generated from a cascade of transformations of the input signal on the retina. As discussed below, the feedforward and hierarchical combinations of relatively simple signal transformations (e.g., Gaussian-like and max-like) can produce a wide range of clutter effects compatible with the experimental data.

Consider the case where a maximum operation is followed by a Gaussian operation, as illustrated in Fig. 5-5. Suppose that the output neuron receives inputs from two afferent neurons (represented by two axes) and that its response is largest for the input pattern generated by the stimulus X. The strength of the neural response is indicated by the gray-level in the figure. A different stimulus Y produces a different input pattern that deviates from the optimal, and thus the neural response will be smaller. When both stimuli, $X+Y$, are presented simultaneously, the afferent neural response is assumed to be determined by the maximum operation. Then, the resulting afferent activation pattern is closer to the optimal pattern, and the output response to the clutter condition is somewhere between the responses to the two isolated stimuli.[3,4]

[3]This simple case requires certain assumptions. For example, it was assumed that the afferent response to the clutter was equal to the maximum of the responses to the single, isolated stimuli. This assumption will not in general be true. Even a neuron computing an exact maximum operation can not guarantee such a response property, because the neural computation operates on the afferent responses, not directly on the individual stimulus. However, if the individual stimuli activate different sets of afferent neurons, whose receptive fields are not overlapping, then such an assumption may hold. There are also some physiological evidences for a max-like response property [Gawne and Martin, 2002; Lampl et al., 2004]. It is important to make the distinction between the max-like operation acting on the neural responses (as proposed by the canonical neural circuitry) and the max-like response behavior to multiple stimuli (as observed in physiological experiments). In Fig. 5-5, however, for the simplicity and illustration purpose, it is assumed that the max-like operation on the neural responses has generated the max-like behavior on the individual stimuli. Similar clutter effects will in general be observed even when this assumption is strictly not valid, as the simulation results show.

[4]Even from this simple illustration, it is clear that there are multiple factors influencing the clutter effect. For example, the afferent activity may be affected by the clutter condition differently, depending on (a) the

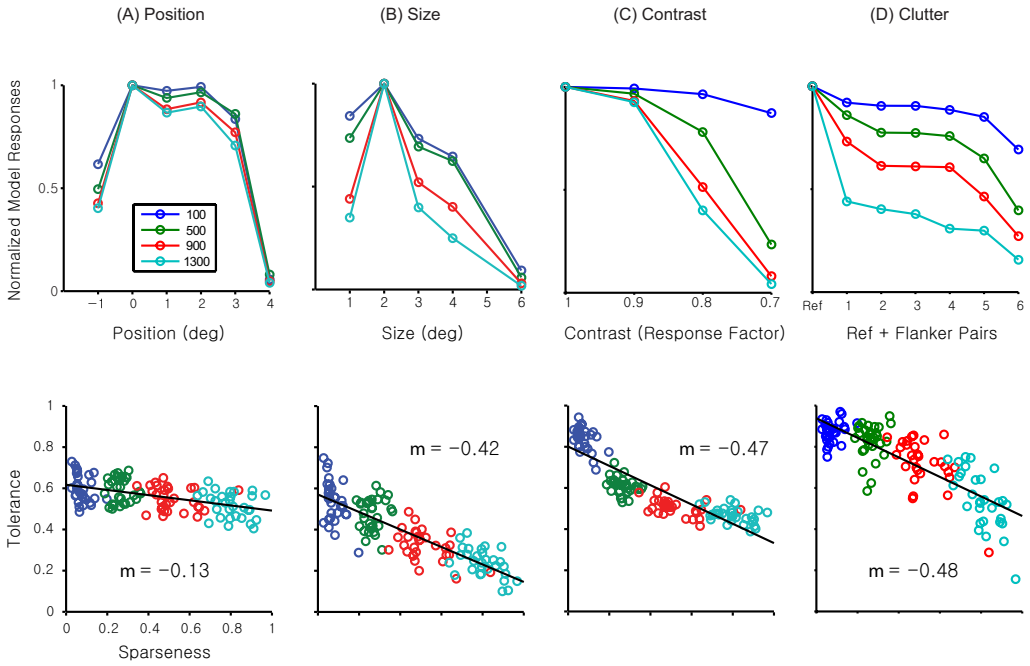


Figure 5-4: In the top panels, the tuning curves of 4 model IT units (with different number of afferent units as indicated in the legend of the first panel) to four different stimulus transformations (position, size, contrast, and clutter) are shown. In the bottom panels, the sparseness and the tolerance indices to these four transformations are plotted for a population of model units (a total of 120 model units with 4 different numbers of afferent units). As discussed in the text, the units with more afferents inputs are more selective and less tolerant, resulting in an inverse relationship, similar to that observed in the recorded IT neuronal population. In (A) and (B), because of the maximum-like operations across different positions and scales in the lower layers of the model, the responses of the model IT units show some degree of tolerance to position and scale changes. However, it should be noted that in general the presence of OR-like (e.g., maximum) operations does not guarantee perfect invariance across position and size changes, because these tolerance operations are applied in a hierarchical manner, interleaved by AND-like (e.g., Gaussian) operations. In the model, a hierarchy of OR-like operations followed by AND-like operations is not equivalent to a global OR operation at the top level. Additionally, pooling over only a finite number of positions and sizes over which the model units perform the OR-like operations contribute to the imperfect invariance of the model. Since model IT units perform a Gaussian-like tuning operation over their afferents, those IT units that are more selective (e.g., they receive more inputs) will still be more sensitive to the (not completely invariant) activation of their afferents due to position and size changes. In summary, the model IT units that were more selective (because they received more afferents from the previous layer) were less tolerant to non-optimal activation of their afferent units produced by the changes in the position, size and contrast (A, B, and C), and by the addition of the cluttering object (D). Hence the trade-off between selectivity and tolerance is observed in all cases, in good agreements with [Zoccolan et al., 2007].

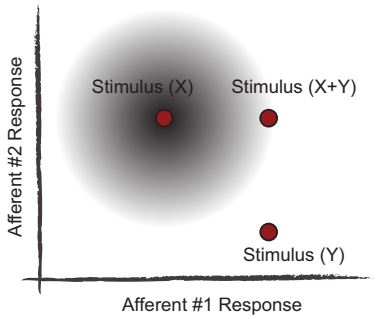


Figure 5-5: A model AIT neuron performs a Gaussian-like tuning operation within the input space of the afferent units. The response is maximal when the activation of the afferents matches the optimal pattern for the neuron, and if the afferent activation deviates from the optimal pattern, the response is small. In this 2-dimensional input example, a neuron tuned to the stimulus X will show a suppressed response if X is paired with another stimulus Y. The afferent activity to the clutter condition lies somewhere between the activities due to isolated X and Y.

Most physiologically observed clutter effects are typically suppressive and rarely facilitatory (i.e., $R(X+Y) < \max(R(X), R(Y))$, where $R(\cdot)$ is the response function of a neuron to a stimulus), and the clutter response is typically greater than the minimum response (i.e., $R(X+Y) > \min(R(X), R(Y))$) [Zoccolan et al., 2005; Reynolds et al., 1999]. According to the simplified model shown in Fig. 5-5, one key mechanism is the tuning operation, by which the neuron responds strongly to the target object (X). When a different object (Y) is presented, the response of the neuron decreases significantly. When the target and the cluttering objects ($X+Y$) are presented together, the neural response decreases less because the corresponding afferent activity is similar to the optimal one, because of the other key mechanism, the maximum-like operation.

Going beyond the schematic illustration shown in Fig. 5-5, Fig. 5-6A shows the responses of the afferent neurons in the full model (i.e., C2b units in Fig. 3-1) to two different objects presented in isolation and in clutter condition. The response of a model IT neuron that receives inputs from those afferents is shown in Fig. 5-6B. The clutter effect from each stimulus pair is summarized with the clutter index in Fig. 5-6C. The index value less than 1 indicates that the effect is suppressive. Note that the afferent responses do not always follow the max-like behavior as in Fig. 5-5, even though the afferent neurons are performing

type of operation that the afferent neurons are performing, (b) the receptive field size and properties of the afferent neurons (e.g., if the receptive field of the afferent neuron is large, the clutter stimulus may fall into the receptive field and change its response) and (c) selectivity and tolerance properties of the afferent neurons (e.g., if the afferent neurons themselves are tolerant to the clutter and do not change their responses, there will not be noticeable clutter effect in the efferent neuron). The nature of the efferent neuron also plays an important role, depending on (d) the sensitivity of the neural response (e.g., if the efferent neuron is highly sensitive to the change of the afferent activation patterns, the clutter stimulus will produce a large change in the neural response) and (e) the operation type (e.g., the Gaussian-like operation produces a different clutter response, compared to the maximum or summation-like operations). The experimentally observed clutter effects will certainly involve multiple factors, and they will be in general difficult to distinguish or isolate. For instance, an average-like interaction, instead of the maximum, in Fig. 5-5 would yield similar responses to the clutter. Nevertheless, the presented simulation results indicate that a hierarchical combinations of relatively simple neural operations is capable of producing different clutter effects observed in physiological data, and in some cases, the results may even be unintuitive (e.g., maximum operation does not necessarily imply a complete clutter tolerance).

maximum-like operation. However, the suppressive clutter effects are still observed, in agreement with the neural data.

5.3.4 Possible Mechanism of the Tradeoff Behavior

From a computational perspective, being able to recognize an object in clutter would require two conflicting requirements. On one hand, the recognition system needs to be selective for a target object to be recognized, but on the other hand, the system needs to be tolerant or invariant to changes in the raw inputs caused by the cluttering objects. If the recognition system is highly selective and sensitive for a particular input pattern (e.g., a target object appearing in isolation), the system may not be tolerant to the clutter, as the cluttering object changes the input to the system. Therefore, the more selective the system is, the more perturbed its response would be due to the clutter. Likewise, if the recognition system is tolerant to the clutter, it is expected to be less selective. Such a tradeoff between selectivity and invariance is expected, and indeed observed in the model (Fig. 5-4) and in physiology [Zoccolan et al., 2007], for different image transformations (e.g., position, size and contrast changes) that cause any significant change in the input to the recognition system. Therefore, a tradeoff behavior would be generally observable in any model of object recognition that employs the tuning or template-matching mechanisms.

As schematically shown in Fig. 5-7, having a population of model IT units with a variable number of afferents produced wide ranges of selectivity, tolerance, and the tradeoff between these two properties, in agreement with the experimental findings. The reason is that, as the number of afferents increases, it becomes more difficult for an arbitrary visual stimulus to produce a matching afferent activity to the stored optimal pattern. As a consequence, the selectivity of the unit increases. At the same time, the tolerance of the unit decreases, since it becomes easier for some stimulus transformations (such as adding arbitrary flanker (clutter) stimuli or changing the position, size and contrast of the preferred stimulus) to produce large deviations from the optimal input pattern. This assumption that different neurons may have different number of afferents has an interesting implication. The variation in the number of afferents may be due to a random, noisy process, but a more interesting possibility is that there is an active feature selection process, where connections are made with useful afferent neurons, while less useful inputs are pruned out. During this process, some neurons would end up with more (or less) afferent inputs than others.

The essential mechanism of the tradeoff behavior in the model can be illustrated by a simple toy simulation, as shown in Fig. 5-8. The minimal component in this toy model is the template-matching tuning operation, whose sensitivity can be altered (e.g., by changing the input dimensionality). As in Fig. 5-4, the selectivity and invariance are traded off across the population of toy model units. The simulation results from the full model, as shown in Fig. 5-4, show that tradeoff between selectivity and tolerance is observed in a more complex hierarchical neural network that approximates the architecture of the ventral visual pathway and even contains OR-like (i.e., max-like) operations to build invariance for size and position changes. Similar results are obtained when different model parameters (other than the number of afferents) are varied to obtain a wide range of selectivity and tolerance properties (see Appendix D.2).

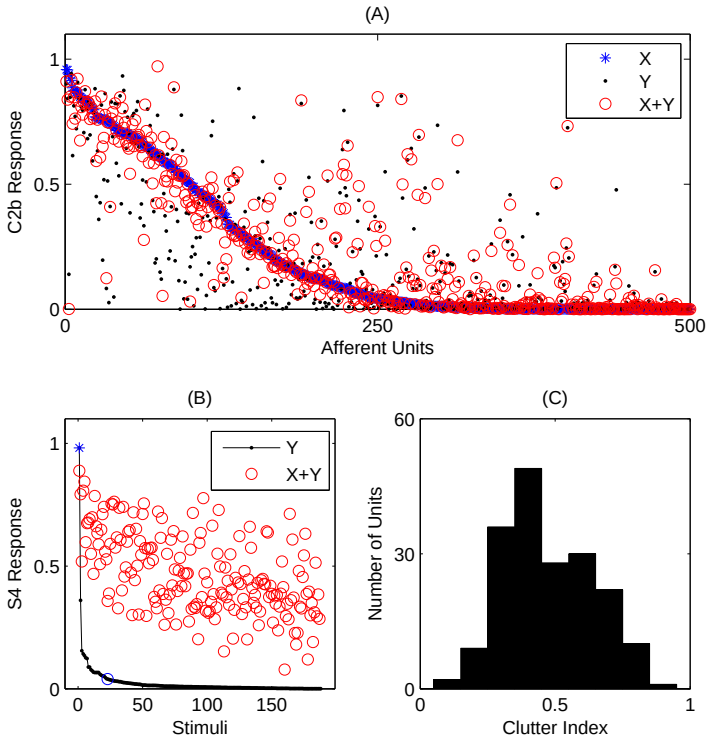


Figure 5-6: This figure shows an example of the clutter effect in the model. A computational model has the advantage of being able to record simultaneously the input and output activities. (A) shows the responses of 500 random afferent neurons to three different stimuli: an object X and Y shown in isolation and in clutter. Note that the afferent activation due to X is more similar to that of $X+Y$ than Y (cf. Fig. 5-5). The interaction at the afferent level is not exactly the maximum, because a more biologically plausible maximum-like operation, soft-max with the exponent 3 in the numerator, was used and because the clutter stimulus can also change the response of the afferent units themselves. The correlation coefficient between X and Y was 0.71, whereas it was 0.85 and 0.88 between X and $X+Y$ and between Y and $X+Y$. Such a trend is found for all combinations with different stimuli, and therefore, a neuron that is selective for the stimulus X produces an intermediate response when X and Y are presented together, as shown in (B) and (C). The target stimulus is indicated by blue * in (A) and (B). The afferents themselves are tuned to the features found in the stimulus objects. Other types of afferent units (e.g., tuned to features found in natural images) generate typically very low responses to the stimulus set, and hence not used. The tuning operation at the S4 layer in the model is a Gaussian-like tuning function based on the normalization circuit, Eq. 2.3.

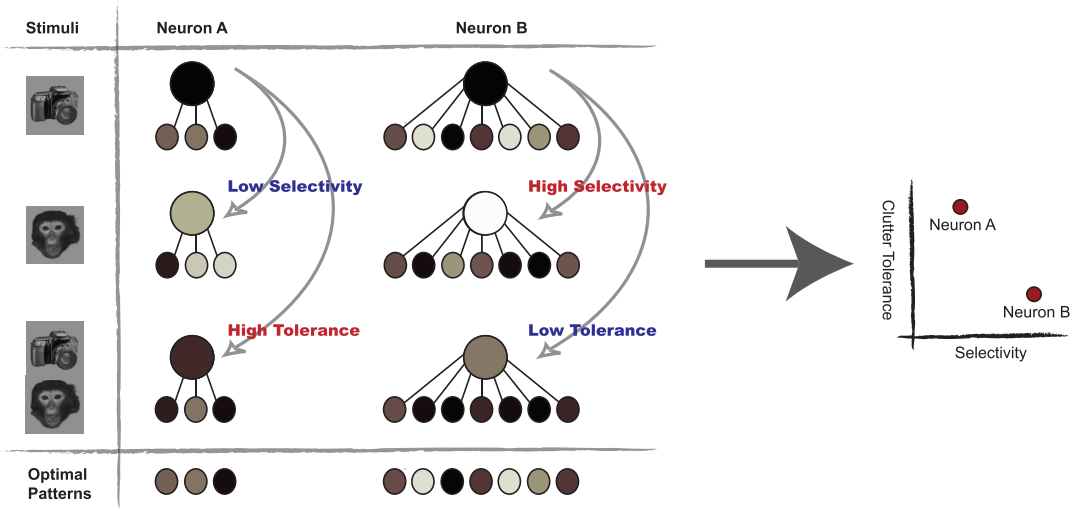


Figure 5-7: Consider two different model IT neurons, A and B, tuned to a particular object, an image of a camera in this example. Each circle denotes a neuron, whose response is indicated by the color scale (darker color means higher response). Each neuron is selective for a particular input pattern (shown at the bottom), corresponding to the response of the afferent V4 or PIT neurons. Both neurons A and B, then, produce high responses to an image of a camera. When the stimulus is changed to a different object (monkey face) or to a clutter condition (camera plus monkey face), the activation pattern of the afferent layer changes accordingly. The neuron A, which is less sensitive to the change in the input (because it receives smaller input changes), shows less reduction in the response than the more sensitive neuron B. The amount of response reduction, whether it is due to the change in the identity of the stimulus or due to the clutter, depends on the sensitivity of the neuron, and. therefore, the selectivity and tolerance properties will be traded off, as shown on the right.

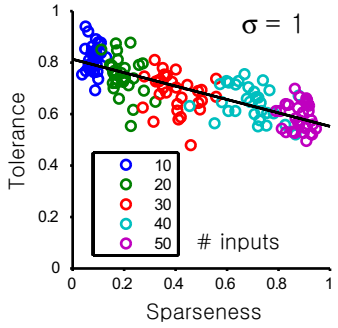


Figure 5-8: Toy simulation. In this simplified setup, a stimulus input is assumed to create a random afferent activation pattern, and the efferent neural response is modeled by the output of a multi-dimensional Gaussian function with fixed σ, equal to 1. The selectivity is measured by the response of the Gaussian function to 200 randomly chosen afferent activities. In the clutter condition, a randomly chosen subset of the afferent units takes the response values to the reference object, while the rest takes the response to the cluttering object (the maximum or averaging interactions yield the similar results). Different numbers of afferent inputs are considered, as indicated by the legend. Similar to the results in Fig. 5-4, this toy simulation produces qualitatively similar tradeoff between selectivity and tolerance properties across the toy model units with different number of afferent units.

5.3.5 Summary

In this section, it was shown that a population of IT-like S4 units in the model of the ventral pathway (Fig. 3-1) shows the similar clutter effects and tradeoff of selectivity and tolerance properties, as observed in IT [Zoccolan et al., 2007]. The model provides a straightforward explanation for the experimental findings, based on a hierarchical, feedforward combination of the Gaussian-like and max-like neural operations.

Chapter 6

Discussion

6.1 Summary and Contributions

In this thesis, I have described a computational model of the ventral pathway [Riesenhuber and Poggio, 1999b; Serre et al., 2005], which is a member of a general class of feedforward, hierarchical models for object recognition [Fukushima et al., 1983; Hubel and Wiesel, 1965; Mel, 1997; Wallis and Rolls, 1997]. Its instantiation has been extended towards a more biologically plausible one, by considering potential neural circuitry and comparing with the neurophysiological data. The contributions of this thesis are:

- In Chapter 2, a canonical neural circuit, involving biologically feasible and observed mechanisms of divisive normalization and polynomial nonlinearities, was studied. This circuit is shown to be capable of generating a variety of neural response patterns, including the Gaussian-like tuning and the maximum-like nonlinear behaviors.

- The current work has demonstrated the plausibility of an idea for a universal elementary unit of computation [Mountcastle, 2003; Douglas et al., 1989], by applying the proposed canonical circuit repeatedly to the full model of the ventral pathway and showing that, as a case in point, it can provide the necessary computations for the selective and invariant neural representations and object recognition performance.

- In Chapter 4, the biological plausibility of the model, especially in the intermediate level (i.e., S2 and C2 units), has been demonstrated by modeling the shape selectivity and translation invariance properties of V4 neurons in response to the bars, gratings, and contour feature stimuli [Desimone and Schein, 1987; Gallant et al., 1996; Pasupathy and Connor, 2001].

- In Chapter 5, the biological plausibility of the model, especially in the highest level of the hierarchy, has been demonstrated by three different comparisons with the neural data from inferior temporal cortex. First, the IT-like model units were shown to have selective and invariant (to translation, scale, and rotation in depth) properties in agreements with [Logothetis et al., 1995]. Second, it was shown that, using a statistical classifier, the responses of the IT-like model units can be used to accurately and invariantly (to translation and scale) categorize the presented stimuli, in agreement with [Hung et al., 2005]. Third, in agreement with [Zoccolan et al., 2007], the model units were shown to have similar suppressive effects to the clutter and the tradeoff of their selectivity and tolerance properties.

There have been several recent studies of showing strong and competitive object recognition performance of the same model, in comparison with humans and other computer vision algorithms [Mutch and Lowe, 2006; Serre et al., 2005, 2007a,b]. These results indicate that such a feedforward, hierarchical model is not only biologically plausible, but also capable of successfully performing a certain class of complex object recognition tasks (i.e., rapid, at-a-glance recognition).

6.2 Open Questions

There are several open questions that future research needs to address.

1. The operations of the canonical circuit in Chapter 2 has been derived under the assumptions of steady state and exact shunting inhibition. It will be important to study such a circuit using the models of realistic synapses and dynamical properties of the real neurons. There is an already preliminary result indicating that the transient response of a more biophysically-detailed circuit can show the Gaussian-like and max-like behaviors [Knoblich et al., 2007]. It will also be interesting to investigate the cell types and laminar structures that may be associated with a canonical circuitry.

2. As indicated in Chapter 2, there is a large body of modeling works using divisive normalization and polynomial nonlinearities, and it is possible that different nonlinear response properties of the cortical neurons, as observed in physiological experiments, may have been generated by such mechanisms. The future study will need to address the similarities and differences among these models. For example, in many divisive normalization models, the polynomial nonlinearities are assumed to be fixed (usually with the exponent of 2) and to be matched at the numerator and the denominator. In the current study, a more general situation has been assumed by allowing different nonlinearities in the divisive normalization terms.

3. The simulation results in Chapter 5 have shown that, even with approximations to the Gaussian and maximum operations, the response of the model units at the top of the hierarchy was selective and invariant enough to support selective and invariant (e.g., to translation and scale) object recognition. It will be important to explore the limits of the model, in terms of its robustness. For example, some pilot studies have shown that quantizing the response of each model unit to just 4 levels, instead of having a continuous analog value, does not change the results shown in Chapter 5. However, the robustness of the model, in the presence of noise, neuronal deaths, synaptic failures, etc., has to be studied more systematically under more realistic and difficult tasks like [Serre et al., 2007b].

4. The current study has dealt with the static behavior of the feedforward, hierarchical network, omitting the dynamical properties of the neurons and their networks. Such an omission was possible by focusing on a more manageable and more narrowly-defined problem of rapid object recognition process. However, the future study needs to incorporate the neural dynamics that will play an important role through synchrony, adaptation, and response modulations coming from the feedback and recurrent connections.

5. The global structure of the system, based on the anatomy and physiology of the ventral pathway, has been designed to be feedforward and hierarchical. Interestingly,

other cortical pathways (e.g., the dorsal visual pathway [Rust et al., 2006] and the auditory cortex) are thought to share similar architecture and perform similar computations. Such similarities across different cortical areas may be natural, because they are all performing object recognition in one way or another (i.e., recognizing a visual motion or an auditory object). It will be interesting to further explore more detailed correspondences and differences between the cortical areas and investigate the constraints (both evolutionary and computational) that give rise to such a global structure. Also, it is an open question how the global network architecture (e.g., hierarchical vs. non-hierarchical) affects the efficiency of the system.

6. The visual recognition system must learn the temporal and spatial correlations from the visual experiences. For example, to recognize a specific face regardless of its poses, the visual system needs to learn, either explicitly or implicitly, how the constituent features of a face, such as the nose and eyes, are transformed and related at various rotations, sizes, and positions during the natural viewing experiences. The proposed canonical neural circuit, which may operate throughout the cortex, may be used to understand the cortical learning mechanisms in a biophysically plausible and concrete context. For example, the strengths of the synaptic weights (and possibly other nonlinearities in the neural circuit) may be adjusted according to the experimentally discovered plasticity rules, and the important statistical properties of the environment may be learned. Then, the resulting neural properties (e.g., their specificity, invariance, sparseness etc.) may be compared to those of a real neuronal population. The plausibility of different learning rules can also be assessed by examining the overall performance of the system before and after applying the learning algorithms.

7. Everyday experiences and psychophysics experiments indicate that attention plays an important role in perception, significantly improving the recognition performance in demanding tasks (e.g., finding a particular person in a crowd). It is widely believed that feedback and back-projecting connections, which are abundantly found in the cortex, mediate the attentional processes [Lee and Mumford, 2003], but their neural mechanisms are unknown. The current model in Fig. 3-1 was originally designed to account for the rapid object recognition process in the absence of attention, but it may be extended to start asking questions about the attentional mechanisms. In particular, the biased competition model [Reynolds et al., 1999], which describes the effects of attention in the visual cortex, involves the similar divisive normalization as the canonical circuit in Chapter 2. The physiological attentional effects can be readily obtained by biasing certain synaptic weights through top-down feedback connections, and such feedback circuitry may be replicated throughout the entire hierarchical layers of the model. It will be interesting to study how such top-down connections may be utilized in the cortex and how they may impact the overall recognition performance of the model. The research paradigm, of investigating the response properties at the neuronal level and verifying the overall performance of the full model at the system level, has proven effective in establishing the current feedforward model of the ventral pathway, and it may again prove effective in understanding the cortical feedback mechanisms.

8. A biologically-inspired model can reveal how the cortex performs difficult visual tasks and at the same time provide new insights and ideas for engineering a more effective vision system, bridging the neuroscience and computer science. The current model in

Fig. 3-1 is itself a computer vision system, which can be run on a personal computer and has outperformed some of the computer vision algorithms in several benchmark object recognition tasks [Serre et al., 2007b]. Hence, the model has a potential for real-world applications, with some enhancements for its speed and performances. Since all the operations in the model are computed by the same canonical neural circuitry, the improvements in the software and the specialized hardware implementations of this core circuit may provide the needed enhancements to be a more feasible real-world system. The investigation into the effective learning algorithms (for training the model) and feedback mechanisms will also contribute to improving the performance of the model as a better computer vision system.

The current thesis has tried to generate a deeper understanding of the computational principles of the complex neural networks in the visual cortex, opening up a number of new questions. Successfully answering them will require continuously close collaborations and interactions between the models and the experiments, in order to develop critical new experiments and ideas about the representations and information processing in the cortex and hopefully to arrive at a coherent and unifying theory, capable of integrating a larger body of experimental results, producing testable hypotheses, and revealing how different computational mechanisms work together to comprise an exquisitely power system.

Appendix A

More on the Model

A.1 Software Implementation of the Model

The overall architecture of the model is hierarchical and feedforward, reflecting (1) the hierarchical organization of the ventral pathway, (2) the rapid recognition performance that does not leave much time for feedback or recurrent processes (hence, focusing on the first few hundred milliseconds of visual processing without attention), and (3) gradual build-up of more complex selectivity and invariance along the hierarchy in order to meet those two requirements for robust object recognition (by repeatedly employing Gaussian-like and max-like operations in an interleaved fashion throughout the hierarchy).

The overall architecture of the model is implemented with three components: Neurons, synaptic weights, and operations. Each model unit (a "neuron") is connected to a number of afferent or input units (i.e., the "presynaptic neurons"). Either Gaussian-like or max-like operation (depending on the functional role of each model unit) is performed over the afferent responses (a continuous, analog value between 0 and 1), using a weighted summation (with synapses), polynomial nonlinearity, and a divisive normalization (with inhibitory connections), which can be implemented by a canonical neural circuitry. Given the hierarchical and feedforward nature of the model, these three components determine the output of the model to any stimuli.

- **Neurons:** For a given stimulus, the response of each neuron in the hierarchy is stored in a 3-d matrix, where the first two dimensions correspond to the spatial dimensions (horizontal and vertical coordinates, x and y) of the receptive field of a neuron. The third dimension is for the type of selectivity of the neuron (i.e., all the neurons in the same third dimension have the same set of synaptic weights with respect to their afferent units, and their receptive fields cover different locations within the visual field). For example, the S1 units (analogous to the orientation-selective V1 cells) make up 2-d maps with different orientation selectivities. Hence, there are three intuitive dimensions to represent the collection of the S1 units (two spatial, retinotopic dimensions and one dimension for the orientation selectivity). In the higher levels in the hierarchy, there are many different types of selectivity (no longer simple as orientations) and the third dimension may be quite large. Furthermore, as the size of the receptive field grows along the hierarchy, the top layer does not have retinotopic organization (i.e., the receptive field covers the entire visual field) and, hence, the first two dimensions are just 1's (i.e., a 1x1xN matrix). Responses to different stimuli are stored in different 3-d matrices.

The receptive field sizes (i.e., scales or resolutions) of different model units even within the same level may be different. The model units belonging to different scales are stored in different 3-d matrices with a "sN" tag, where N is the scale index. Hence, the response of the model units within each level are stored in a structure of multiple 3-d matrices, `r.sN`.

- **Synaptic weights:** For a given model unit, its connectivity with the afferent model units is specified as a linear synaptic weights and stored in a 4-d matrix. The first two dimensions of this matrix are for the spatial (x and y) dimension of the afferent neurons, and the third dimension is for different selectivities of the afferent neurons. Along the fourth dimension, different connectivity types are stacked. In a sense, the responses of the efferent neurons are determined by the 3-d convolution-like operations between a filter (specified by the set of synaptic weights) and the responses of the afferent neurons. Therefore, the size of the third dimension of the 4-d synaptic weight matrix is required to match the size of the third dimension of the 3-d response matrix of the afferent layer. The third dimension of response matrix for the efferent layer (postsynaptic neurons) will have the same size as the fourth dimension of the weight matrix. Such 4-d representation of the synaptic weights is a natural convention to accompany the 3-d representation of the neural responses.

  Two different sets of connectivity (f1 and f2) can be specified in the software, where f1 specifies the direct synaptic weights to the efferent neuron and f2 specifies the indirect synaptic weights (potentially through an inhibitory pool cell for divisive normalization). In principle, f1 and f2 may involve a different set of afferent units (e.g., the afferent units specified by f2 can have collectively larger receptive field and may form a bigger non-classical receptive field). In practice (in the current implementation), f2 has the same size as f1 and its values are fixed at 1.

  These 4-d matrices are in general sparse. For example, there may be thousands of different neuron types in the afferent layer (third dimension), yet the efferent neuron may make synaptic connections with only a small fraction from a pool of all available afferent neurons. Hence, in the implementation, the 4-d weight matrices are stored in a sparse format, so that the weighted sum operations are performed for the nonzero weights only, saving both processing time and memory. This packaging of the 4-d matrices is performed by `weights_package.m` function, which produces the following fields from two 4-d matrices (f1 and f2), for a given scale "sN":

  `f.sN.f1`: nonzero weights within f1
  `f.sN.f2`: nonzero weights within f2
  `f.sN.i1`: index for the 1st dimension
  `f.sN.i2`: index for the 2nd dimension
  `f.sN.i3`: index for the 3rd dimension
  `f.sN.size`: size of f1 and f2 (4 numbers for 4 dimensions)
  `f.sN.shift`: sampling of the efferent neurons

- **Operations:** There are two principle operations in the model for achieving selectivity and invariance. In the original implementation of the model by Riesenhuber and Poggio (1999), these two operations were the multi-dimensional Gaussian function and the maximum operation. They can actually be approximated by a simple, biologically plausible canonical neural circuitry that computes weighted summation, polynomial

nonlinearity, and divisive normalization (possibly through a shunting inhibition by a pooling neuron).

The mathematical expression takes the following form (pseudo-code):

```
y = sum(f1.*(x.^p)) / (k+sum(f2.*(x.^q)).^r);
/* weighted sum, polynomial nonlinearity, divisive normalization */
```

Depending on the values of `p`, `q`, and `r`, the output `y` can be a Gaussian-like tuning operation that peaks around some particular input pattern, or a max-like operation (e.g., softmax) that provides selection and invariance.

The computationally intensive part of the model is the weighted summations in the numerator and the denominator. In this implementation (for the lack of massively parallel computer like the cortex), they are computed by the mex-C files. Taking advantage of the sparse nature of f1 and f2, these codes perform weighted sums only for the nonzero weights.

```
[y1, y2] = do_sp2 (x.^p, x.^q, f1, f2, ...);
       y = y1./(k+y2.^r);
       /* where y1 = sum(f1.*(x.^p)) and y2 = sum(f2.*(x.^q)) */
```

A.2 Receptive Fields

Due to the feedforward nature of the model, the receptive field size of a model unit is completely determined by a few parameters of the current and the previous layers. See Fig. A-1.

The receptive field sizes of the model units in the hierarchy are recurrently defined by the following relationship:

$$r'_n = s'_{n-1} \cdot (g_n - 1) + r'_{n-1} \quad \text{(A.1)}$$

$$s'_n = s_n \cdot s'_{n-1} \quad \text{(A.2)}$$

The variable r'_n denotes the receptive field size of a unit in the current (n-th) layer. s'_n indicates how far apart the neighboring units are in terms of the visual field. Thus, s'_n is related to the sampling, or the inverse of the coverage density, or the inverse of the "overlap" between the receptive fields of the neighboring units. For simplicity, a uniform square grid coverage is assumed. The $'$ above r and s indicates that those values pertain to the actual image plane and are derived from other parameters that are more relevant within each layer. g_n indicates the grid size from which a current layer unit receives the input. Thus, this value determines the pooling range of a model unit. Finally, s (without $'$) indicates the sampling for the current layer, in terms of the grid on the previous layer.

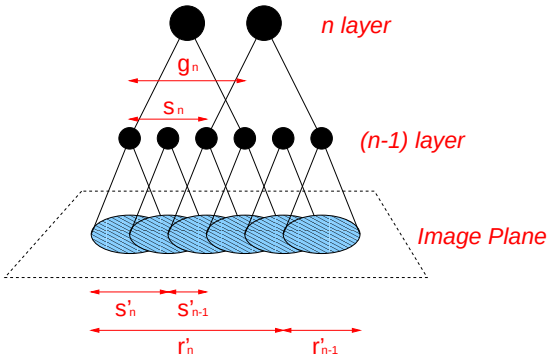


Figure A-1: Schematic diagram of how different model parameters are related and lead to the increasing receptive field sizes.

Appendix B

More on the Canonical Circuit

B.1 Divisive Normalization with Shunting Inhibition

Fig. 2-1A shows a plausible neural circuit that computes a divisive normalization. Fig. B-1 is an equivalent circuit diagram of the cellular membrane of the output neuron y in Fig. 2-1A, which receives the inputs from the neurons x_1, x_2 and x_3, along with the shunting inhibition from a pool cell. The membrane is assumed to have the capacitance C, the variable conductances g's and the equilibrium potentials E's. The subscript o (E_o, g_o) is for the leak ionic channels, e (E_e, g_e) is for the excitatory inputs, and i (E_i, g_i) is for the shunting inhibition. The magnitude of the conductances (g_e and g_i) depends on the level of their presynaptic activities.

Suppose that the transformation of a signal from a presynaptic to a postsynaptic neuron is described by the polynomial nonlinearity with the exponent p, q, or r as in Fig. 2-1A. Then, we may identify

$$g_e \propto \sum_{i=1}^{n} w_i {x_i}^p, \tag{B.1}$$

$$g_i \propto \left(\sum_{i=1}^{n} {x_i}^q \right)^r. \tag{B.2}$$

The membrane potential evolves according to

$$-C\frac{dV}{dt} = g_e(V - E_e) + g_i(V - E_i) + g_o(V - E_o). \tag{B.3}$$

At the steady state,

$$V = \frac{g_e E_e + g_i E_i + g_o E_o}{g_o + g_e + g_i}. \tag{B.4}$$

Now assume that $V_{rest} = E_i = 0$ (i.e., shunting inhibition), $g_e << g_i$ (based on experimental evidences like [Borg-Graham et al., 1998]), and the leak current $g_o E_o$ is not significant compared to other inputs $g_e E_e$. Then,

$$V \sim \frac{g_e E_e}{g_o + g_i}. \tag{B.5}$$

Next, assume that the membrane potential V and the activity of the output neuron y are

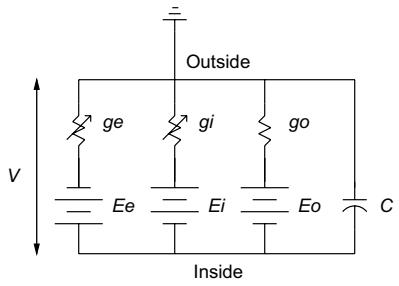


Figure B-1: A circuit diagram of the biophysics of a single cell

proportional. Thus, using Eq. B.1 and B.2, we arrive at Eq. 2.3 in Table 2.1, with positive and negative weights (depolarizing and hyperpolarizing inputs) in general. For a similar discussion, see [Torre and Poggio, 1978; Reichardt et al., 1983; Carandini and Heeger, 1994], but also [Holt and Koch, 1997].

Note that even without explicitly assuming shunting inhibition (i.e., $g_i = 0$), Eq. B.4 contains divisive normalization by the input conductance g_e, which can realize similar gain control mechanism [Borst et al., 1995]. When g_o is much greater than g_e, the overall operation will be close to the weighted summation without normalization.

B.2 Optimal Input Pattern for Divisive Normalization

A straightforward calculus on Eq. 2.3 shows that the peak of the normalized weighted summation operation occurs at

$$x_{oi}{}^{q-p} = \frac{p}{qr} \frac{k + \left(\sum_{j=1}^{n} x_{oj}{}^{q} \right)^{r}}{\left(\sum_{j=1}^{n} w_j \; x_{oj}{}^{p} \right) \left(\sum_{j=1}^{n} x_{oj}{}^{q} \right)^{r-1}} \cdot w_i \tag{B.6}$$

$$\equiv \gamma \cdot w_i. \tag{B.7}$$

The above expression can be interpreted as a tuning around x_{oi} that is proportional to w_i (although, due to the exponent $q - p$, the proportionality is not strictly linear).

For consistency, it is required that

$$\gamma = \frac{p}{qr} \frac{k + \gamma^{qr/(q-p)} \left(\sum_{j=1}^{n} w_j{}^{q/(q-p)} \right)^{r}}{\gamma^{p/(q-p)+q(r-1)/(q-p)} \left(\sum_{j=1}^{n} w_j{}^{p/(q-p)+1} \right) \left(\sum_{j=1}^{n} w_j{}^{q/(q-p)} \right)^{r-1}}. \tag{B.8}$$

For a given set of w's, this condition relates the parameters k and γ. If γ is already determined or fixed (e.g., $\gamma = 1$, as in the simulations), Eq. B.8 will uniquely determine the parameter k. It is unlikely that there is an active neural process that adjusts k for given

w's, as the above derivation suggests. Rather, a fixed constant k (possibly originating from leak currents) and a set of synaptic weights $\vec{w}$ determines the center of tuning $\vec{x}_o$.

Note that the relation between k and γ may not always be uniquely defined. Consider a special case when $(p, q, r) = (1, 2, 0.5)$ (divisive normalization by the L2-norm of $\vec{x}$). Then, Eq. B.8 is satisfied for any γ when $k = 0$. This result confirms the intuition that $\vec{w} \cdot \vec{x}/|\vec{x}|$ has the maximum value as long as these two vectors are parallel. However, in general when $p < qr$, Eq. 2.3 describes tuning around some particular point, $\vec{x}_o$.

B.3 Sigmoid Parameters for Determining the Sharpness of Tuning

The following approximation shows the relationship between the parameters in the sigmoid (Eq. 2.9) and the Gaussian (Eq. 2.1) functions, for determining the sharpness of tuning (see also [Maruyama et al., 1992]).

$$y = e^{(-|\vec{x}-\vec{w}|^2/2\sigma^2)} \tag{B.9}$$

$$\sim \frac{1+S}{1+Se^{(|\vec{x}-\vec{w}|^2/2\sigma^2)}}, \text{for } S \gg 1 \tag{B.10}$$

$$= \frac{1+S}{1+Se^{-\frac{1}{\sigma^2}(\vec{w}\cdot\vec{x}-\frac{|\vec{x}|^2+|\vec{w}|^2}{2})}}. \tag{B.11}$$

We can identify that α in the sigmoid function is related to $1/\sigma^2$. The weighted summation $\vec{w} \cdot \vec{x}$ can replace the computation of the Euclidean distance, and β is related to $(|\vec{x}|^2+|\vec{w}|^2)/2$. Table B.1 shows the sigmoid parameters α and β, which have been obtained by approximating a given Gaussian function using the sigmoidal scaling on the tuned output of the normalization circuit A or B in Fig. 2-1.

As suggested by the above relationship, a narrower Gaussian function (small σ) is approximated with a steep sigmoid function (large α and β). This trend is apparent along each row of the table (especially for the bottom rows). The location of the tuning peak also has a strong influence on the sigmoid parameters, as shown along each column. A Gaussian function centered farther away from the origin tends to require a steeper sigmoid.

For each approximation, 1000 points from a multi-dimensional input space were sampled near the center of tuning. A range of input dimensions (d = 4, 16, 64, 256, and 1024) was considered, but it produced only small variations in the fitted sigmoid parameters, using the normalization of the dimensionality by $\sqrt{d}$. The result reported here is based on the averages of 100 nonlinear fits. Note that a heuristic rule of setting $\alpha = 1/ < y^2 >$ and $\beta =< y >$, where the averages are taken over multiple stimuli, can usually produce a good dynamic range in the output.

B.4 Optimal Templates for the Tuning Operation with L2-norm

Since the scalar product $\vec{x} \cdot \vec{w}$ measures the cosine of the angle between the two vectors, the maximum occurs when they are parallel. Because it is also proportional to the length of each vector, a simple scalar product is not as flexible as the Gaussian function which can have an arbitrary center of tuning. We may assume that both vectors $\vec{x}$ and $\vec{w}$ are normalized

$\vec{w}$ vs. $\sigma/\sqrt{d}$	0.1	0.2	0.3	0.4
0.1	8, 0.61	10±2, 0.33	12±5, 0.20	13±7, 0.12
0.3	25, 0.92	10, 0.76	7, 0.61	6, 0.48
0.5	62, 0.97	18, 0.89	11, 0.79	8, 0.69
0.7	117, 0.98	32, 0.94	17, 0.87	11, 0.80
0.9	192, 0.99	51, 0.96	25, 0.92	16, 0.87

Table B.1: The sigmoid parameters (α, β), for approximating a given Gaussian function with σ and the center $\vec{w}$, are shown. The sigmoid nonlinearity is applied after the tuning operation of Eq. 2.3 in Table 2.1 with $(p, q, r) = (1, 2, 1)$. The Gaussian function has d-dimensional inputs (i.e., d is the number of the afferent neurons) with $\sigma = \sqrt{d}$ times a numerical factor (between 0.1 and 0.4; columns). The $\sqrt{d}$ factor allows the normalization of the Gaussian σ across the different dimensionalities and effectively divides the Euclidean distance by d. The center of tuning was varied along the diagonal in the multi-dimensional input space, such that $\vec{w} = (1, ..., 1)$ times a numerical factor (between 0.1 and 0.9; rows). The variations over the different dimensionality d was usually small (except for the first row, as indicated by ±), because the Gaussian σ was already normalized by the $\sqrt{d}$ factor.

(for example, $\vec{x}$ by the divisive normalization as in circuit A or B in Fig. 2-1, and $\vec{w}$ by Oja's rule [Oja, 1982]), so that only the direction within the input space is relevant.

Because of the normalization, the dimensionality of the normalized scalar product is one less than that of a Gaussian function. With the same number of afferents n, the Gaussian tuning function may be centered at any point in the full multi-dimensional space $\mathbb{R}^n$, whereas the domain of the normalized scalar product is restricted within the hypersphere $\mathbb{S}^n$ or $\mathbb{R}^{n-1}$. A simple and obvious way of avoiding such a limitation is to assume a constant dummy input and to increase the dimensionality of the input vector, as in [Maruyama et al., 1992]. This constant may be the resting activity of the neuron. Then, the normalized scalar product may be tuned to any arbitrary vector $\vec{w}$, just like the Gaussian function. See Fig. B-2A.

Assuming such a constant dummy input (indexed with d in w_d and x_d), the response of the L2-norm neural circuit is given by

$$y = \frac{\sum_{j=1}^{n} w_j \; x_j + w_d \; x_d}{k + \sqrt{\sum_{j=1}^{n} x_j^2 + x_d^2}}, \tag{B.12}$$

which can be viewed as the normalized scalar product in the $(n+1)$ dimension. Then, it is easy to verify that by appropriately choosing w_d and x_d, the maximum response can occur at some arbitrary point, $\vec{x} = \vec{w_o}$.

Let's take the partial derivative:

$$\frac{\partial y}{\partial x_i} = \frac{w_i}{k + \sqrt{\sum_{j=1}^{n} x_j^2 + x_d^2}} - \frac{\sum_{j=1}^{n} w_j \; x_j + w_d \; x_d}{\left(k + \sqrt{\sum_{j=1}^{n} x_j^2 + x_d^2}\right)^2} \cdot \frac{\frac{1}{2} \; 2 \; x_i}{\sqrt{\sum_{j=1}^{n} x_j^2 + x_d^2}}. \tag{B.13}$$

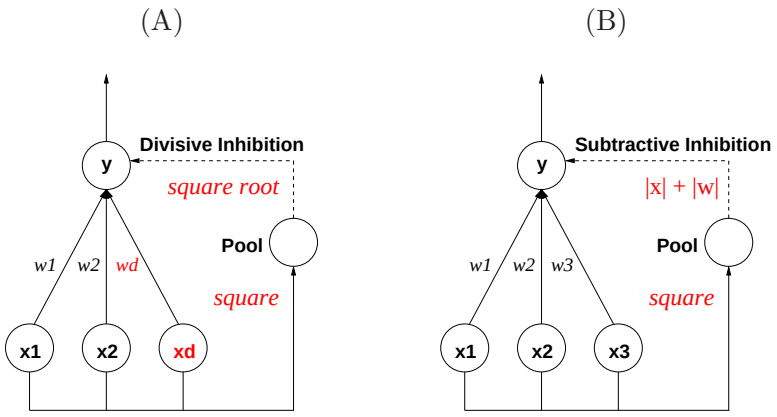


Figure B-2: These neural circuits are two different variations of Fig. 2-1A for performing a Gaussian-like operation. (A) Tuning is achieved by the L2-normalized scalar product. A dummy input (denoted by x_d) may be assumed for a tuning profile centered at any point within the input space. See Appendix B.4. (B) A pooled subtractive inhibition can be used to compute the Euclidean distance between $\vec{x}$ and $\vec{w}$, as discussed in Appendix B.5.

Setting $\frac{\partial y}{\partial x_i} = 0$,

$$0 = w_i \left(k + \sqrt{\sum_{j=1}^{n} x_j^2 + x_d^2} \right) \sqrt{\sum_{j=1}^{n} x_j^2 + x_d^2} - x_i \left(\sum_{j=1}^{n} w_j \; x_j + w_d \; x_d \right). \tag{B.14}$$

Setting $x_i = w_i$ (for all i) and simplifying the expression,

$$w_d = k \sqrt{\frac{\sum_{j=1}^{n} w_j^2}{x_d^2} + 1} + x_d. \tag{B.15}$$

As long as the above condition is met, any arbitrary $\vec{w}$ can serve as an optimal template, and since w_d and x_d can be freely chosen, it is easily satisfied. In particular, set $x_d = 1$ and $w_d = k\sqrt{\sum_{j=1}^{n} w_j^2 + 1} + 1$.

B.5 Relationship Between Circuits

The operations performed by the different neural circuits in Fig. 2-1 are related, especially between the divisive normalization circuits (Eq. 2.3; circuits A and B in Fig. 2-1) and the simpler weighted summation circuit (Eq. 2.4; circuit C).

First, the divisive operation can be approximated by the Taylor series expansion with a subtractive inhibition. For example, $1/(1+\epsilon) \sim 1-\epsilon$, if $\epsilon \ll 1$. In other words, the divisive normalization element ϵ may have the similar effect as the subtractive inhibition.

Second, note that $-(\vec{x}-\vec{w})^2 = \sum_i w_i x_i - \sum_i x_i^2 - \sum_i w_i^2$. Hence, the Euclidean distance measure is composed of the weighted summation and the subtractive elements (sum of the squared inputs). Therefore, if the pool cell in Fig. 2-1A performs a subtractive inhibition instead of divisive shunting inhibition (see Fig. B-2B) and if there is an exponential-like scaling of the output responses, an exact Gaussian tuning can be achieved. In some cases like [Oja, 1982], $\sum_i w_i^2$ can even be assumed constant.

B.6 Examples of Approximation and Learning in Higher Input Dimensions

Considering that the actual neural circuitries in the cortex involve a large number of afferent neurons, it is important to verify that the normalization circuits A and B in Fig. 2-1 are capable of approximating the Gaussian and the maximum operations even when the input dimensionality is large, as shown in Fig. B-3.

Fig. B-4 illustrates that the simple learning mechanism of Eq. 2.11 can be applied to the same circuits with a large input dimension and learn correct synaptic weights. However, the result indicates that it does not scale well with the dimensionality. Such problem is well known for the simple gradient descent algorithms like Eq. 2.11, as finding the global extrema by random drift is too slow in a high dimensional space. Therefore, for a large neural network, more sophisticated and robust learning mechanism is required, or the circuit has to be maintained at some manageable size. The need to keep the number of afferent neurons relatively small for a simple learning mechanism would favor the circuit A or B over circuit C in Fig. 2-1.

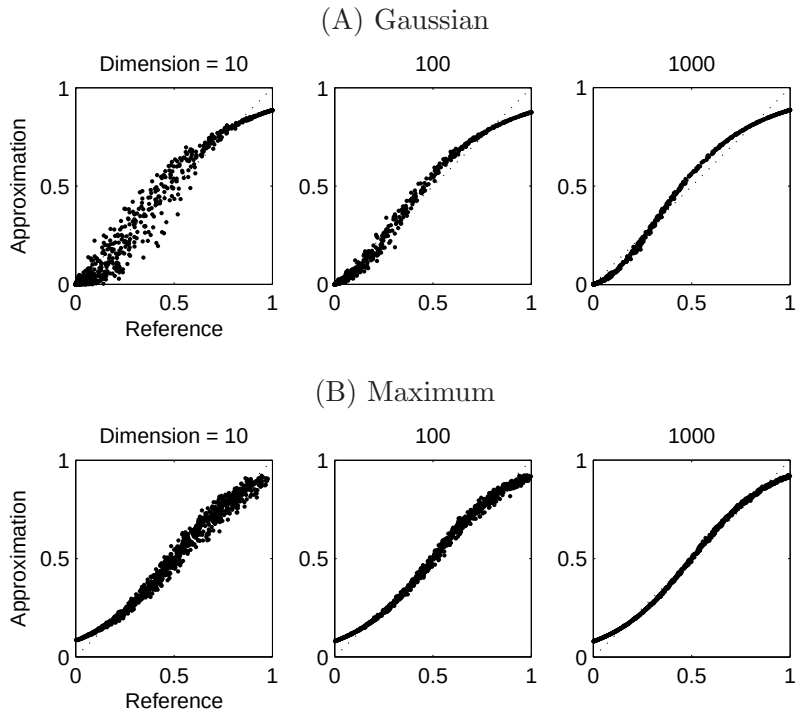


Figure B-3: The Gaussian and maximum operations are approximated by the divisive normalization operation, Eq. 2.3, corresponding to the circuit A or B in Fig. 2-1, with different input dimensions d (10, 100 and 1000). For simplicity, the Gaussian function is centered at 0.5 along each input dimension with $\sigma = 0.2\sqrt{d}$. A nonlinear fitting routine was used to fit the sigmoid parameters, α and β in Eq. 2.9, following the operation of Eq. 2.3 in Table 2.1. In each panel, the ordinate axis is for Eq. 2.3, and the abscissa is for the reference Gaussian or maximum functions; hence, the points along the diagonal line would indicate a perfect fit. For approximating a Gaussian function, the parameters were set at $(p, q, r) = (1, 2, 1)$ with appropriately chosen $\vec{w}$ and k (see Appendix B.2). For max, $(p, q, r) = (3, 2, 1)$ with $k = 0.001$. In order to cover a wider range of the output values, 1000 points in the high dimensional space were pseudo-randomly sampled. For the Gaussian, d-dimensional points are selected from a normal distribution around the center of tuning, along each dimension. For max, uniformly distributed sample points with varying magnitudes (between 0 and 1, also uniformly distributed) are selected. Because of the high dimensionality, sampling each point from a uniform distribution will only cover a limited response range (i.e., a truly random sampling tends to produce only small responses for the Gaussian and only large responses for the maximum). The spread of the points (noise in the approximation) is reduced in the higher dimensions, because the 1000 sample points are less likely to produce the similar responses or to be chosen from a nearby region of the input space.

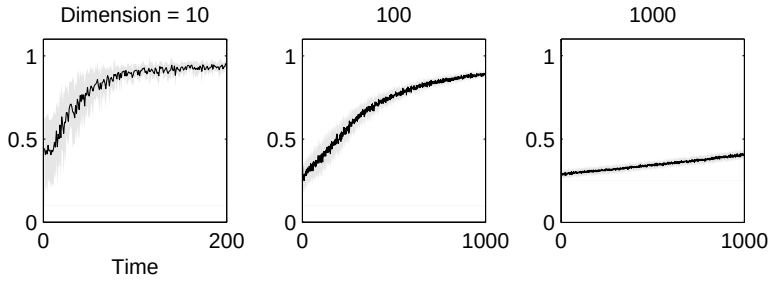


Figure B-4: This figure shows the learning process of Eq. 2.11 in different input dimensions ($d = 10$, 100 and 1000). The average evolutions (the mean and the standard deviations over 10 random initializations in each case) of the output during the learning period are plotted. The neural circuit learns to respond maximally (the maximum response is near 1) to the prescribed pattern of input, while it is exposed to the same input pattern. However, the learning rate is too slow to be practical for the higher input dimensions, as shown in the last panel. Therefore, the neural network will have to either rely on a more sophisticated learning rule or keep a small number of afferent neurons. The input pattern used here is $\vec{x}_o = (0.5, ..., 0.5)$. The parameters in the Gaussian-like operations are again $(p, q, r) = (1, 2, 1)$, corresponding to the case of circuit A or B in Fig. 2-1 and Eq. 2.3 in Table 2.1. The sigmoid scaling of the response, Eq. 2.9, was applied with $(\alpha, \beta) = (20, 0.9)$.

Appendix C

More on the Tuning Properties of Simple S2 Units

In Chapter 4, it was shown that the S2 units (and the efferent C2 units) can have very complex shape selectivity, arising from the nonlinear combinations of the oriented C1 subunits. In this section, a population of very simple, hard-wired S2 units is studied, for an intuitive understanding of how more complex shape selectivity may arise from a combination of simpler ones, using the responses to the bars, Cartesian and non-Cartesian gratings. Although the S2 units do not have as much translation invariance as the C2 units or the V4 neurons do, the selectivity of the S2 units will be compared to that of the V4 neurons, in order not to involve another nonlinear step between the S2 and C2 layers. Note that in the full model of the ventral pathway (Fig. 3-1), the selectivity of the "S" units are learned, in an unsupervised way, from hundreds of natural images, instead of being hard-wired. More detailed version of this study appeared in a technical memo, [Kouh and Riesenhuber, 2003].

C.1 Simple S2 Units

The HMAX model [Riesenhuber and Poggio, 1999b] is a precursor of the current model [Serre et al., 2005], and it is composed of four hierarchical, feedforward layers, labelled as S1, C1, S2, and C2. In this version of the model, each S2 unit combines four adjacent C1 afferents in a spatial 2x2 arrangement, producing a total of 256 (4^4, 4 spatial positions with 4 orientations) different types of S2 units. Each S2 unit performs an Gaussian operation (with fixed σ), and the maximum response is produced when the four C1 afferents generates their own maximum responses. In other words, the response of an S2 unit is determined by the afferent C1 units, which are combined as a product of Gaussians, where each Gaussian is centered at 1:

$$S2 = e^{-\left(\sum_i (C1_i - 1)^2\right)/2\cdot(\sigma_{S2})^2}. \tag{C.1}$$

In the following simulations, the responses of the S1 and S2 units to different sets of stimuli (oriented bars and gratings) are measured with baseline-subtraction, where the baseline was defined to be the response to a null stimulus. The simulation procedures and stimuli were based directly on the physiological studies of the macaque monkeys [De Valois et al., 1982; Desimone and Schein, 1987; Gallant et al., 1996], so that the simulation results could be directly compared with the experimental data.

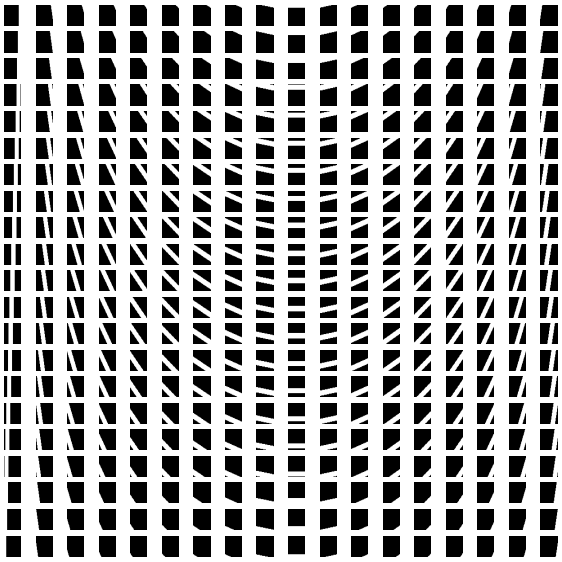


Figure C-1: Examples of bar stimuli at varying orientations and positions. Each square corresponds to the receptive field of a model unit, and the width of the bars shown here is equal to 25% of the receptive field size.

C.2 Methods: Bars and Gratings

The procedures for the orientation selectivity study followed that of [Desimone and Schein, 1987]. The stimuli were the images of bars at varying orientations (0° to 180° at 10° intervals) and widths (1, 5, 10, 15, 20, 25, 30, 50, and 70% of the receptive field size). The bars were always long enough to cover the whole receptive field and presented at different locations across the receptive field, as shown in Fig. C-1.

The orientation tuning curve of a model unit was obtained by finding the preferred width of the bar stimulus and then measuring the maximum (baseline-subtracted) responses over different bar positions at each orientation. The orientation bandwidth was defined as a full width at half maximum with linear interpolation. Fig. C-3 shows the examples of orientation tuning curves.

Again following the convention used in [Desimone and Schein, 1987], the contrast of the bar image was defined as the luminance difference between the bar and the background, divided by the background luminance. Throughout the experiment, the stimulus contrast was fixed at 90% (The results were similar for a wide range of contrasts).

The procedures for the grating selectivity study followed that of [Gallant et al., 1996]. Three classes of gratings (Cartesian, polar, and hyperbolic) were prepared according to the same equations in [Gallant et al., 1996].

The contrast of the grating stimuli was defined by

$$\text{Contrast} = \frac{L_{max} - L_{min}}{L_{max} + L_{min}}. \tag{C.2}$$

The mean value of the grating was set to a nonzero constant, and its amplitude of modulation was adjusted to fit the contrast of 90%.

These gratings were presented within the receptive field of a model unit at varying phases, in steps of 120° and 180° (as in [Gallant et al., 1996]), and the baseline-subtracted

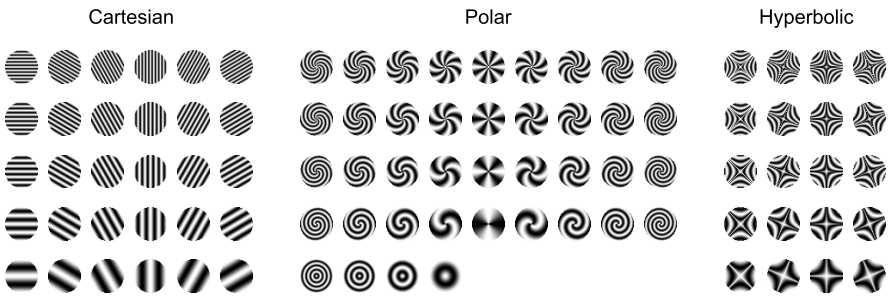


Figure C-2: Grating stimuli (30 Cartesian, 40 polar, and 20 hyperbolic gratings) as used in [Gallant et al., 1996].

	Experiment		Simple S2	
Bandwidths	V1	V4	S1	S2
Median:				
All neurons	42°	75°	39°	77°
Less than 90°	37°	52°	39°	59°
Percentage:				
Narrow (< 30°)	27%	5%	0%	0%
Wide (> 90°)	15%	33%	0%	11%

Table C.1: Summary of the physiological data (V1, V4) and the simulation results (S1, S2). The experimental data were taken from [Desimone and Schein, 1987; De Valois et al., 1982].

maximum responses were calculated.

C.3 Results: Orientation Selectivity

Neurons in visual area V1 exhibit varying degrees of orientation selectivity. The upper left histogram in Fig. C-4 shows the distribution of orientation bandwidth in V1 (from [De Valois et al., 1982]). The median is 42°, while the median of the oriented cells alone (bandwidth < 90°) is 37°. These results are summarized in Table C.1.

In the original HMAX model [Riesenhuber and Poggio, 1999b], each S1 feature was modeled as a difference of Gaussians. However, these features turn out to have an orientation bandwidth much broader (approximately 90°) than found in the experiment [Serre and Riesenhuber, 2004], and the Gabor filters were shown to provide better approximations to the experimental data in V1 [Dayan and Abbott, 2001; Ringach, 2002]. A Gabor filter is defined as

$$G(x,y) = \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right)\frac{\cos(kx-\phi)}{2\pi\sigma_x\sigma_y}. \tag{C.3}$$

By varying σ_x, σ_y, and the wave number k, the properties of the Gabor filter can be adjusted [Serre and Riesenhuber, 2004]. The following parameters are used: Spatial phase $\phi = 0$, so that the peak is centered. Spatial aspect ratio, $\sigma_x/\sigma_y = 0.6$. The extent in the x direction, $\sigma_x = 1/3$ of the receptive field. The wave number $k = 2.1 \cdot 2\pi$. In this neighborhood of k, there are two inhibitory surroundings and one excitatory center. These parameters were chosen to produce a median bandwidth of 39°, close to the median of the V1 bandwidths.

Group	Example	Number	Afferent Configuration
8	||||	4	All 4 in the same orientation.
7	|||/	32	3 in the same orientation, and the other at non-orthogonal orientation.
6	|||-	16	3 in the same orientation, and the other at orthogonal orientation.
5	||//	24	2 in the same orientation, and the other 2 in the same orientation that is non-orthogonal to the first 2.
4	||-/	96	2 in the same orientation, and the other 2 at different and non-orthogonal orientations to each other.
3	||/\	48	2 in the same orientation, and the other 2 at different and orthogonal orientations to each other.
2	||--	12	2 in the same orientation, and the other 2 in the same orientation that is orthogonal to the first 2.
1	-|/\	24	All 4 in different orientations.

Table C.2: 8-class classification scheme for the 256 S2 units. In the *Example* column, the four characters represent the possible orientations of the afferent C1 units. The 2x2 geometric configuration was written as a 1x4 vector for notational convenience. The *Number* column shows the number of S2 units belonging to each class.

A Gabor filter produces an optimal response when the bar stimulus is oriented along the same direction as the filter itself. For a given set of parameters (σ_x, σ_y, and k), the orientation tuning curves of the Gabor filters at different sizes are almost identical to one another. Therefore, in the model, the distribution of the orientation bandwidths in the S1 layer is very sharply peaked around a single value. However, even from this extremely homogeneous S1 population, the feedforward feature combination at the next S2 layer can create a wide variety of model units with different orientation bandwidths.

Moving from V1 to V4, the receptive field size increases, and neurons respond more to the shapes of intermediate complexity [De Valois et al., 1982; Desimone and Schein, 1987; Mahon and De Valois, 2001; Gallant et al., 1996]. In the model, the C1 afferents are combined in a predefined 2x2 arrangement, and, therefore, each S2 unit can be categorized according to its geometric configuration of the four afferents, as shown in Table C.2. Such a classification scheme turns out to be meaningful for characterizing the behavior of these S2 units. For example, each orientation tuning curve in Fig. C-3, typical of each class, shows that the responses to the bar stimuli are determined by how the afferent features are geometrically combined. Some model units, whose afferents are aligned in the same orientations, have very simple unimodal tuning curves, resembling that of an S1 unit (group 8). For others (group 2–7), the tuning curves show multiple peaks at different orientations. Those in group 1, whose afferents are at orthogonal or non-parallel orientations to one another, exhibit little or no orientation tuning.

As a result, the S2 units with similar feature configuration tend to have similar orientation bandwidths, as seen in Fig. C-4. The S2 units in group 6, 7 and 8 have narrow bandwidths around 40°. Group 1 has an extremely broad orientation tuning profile due

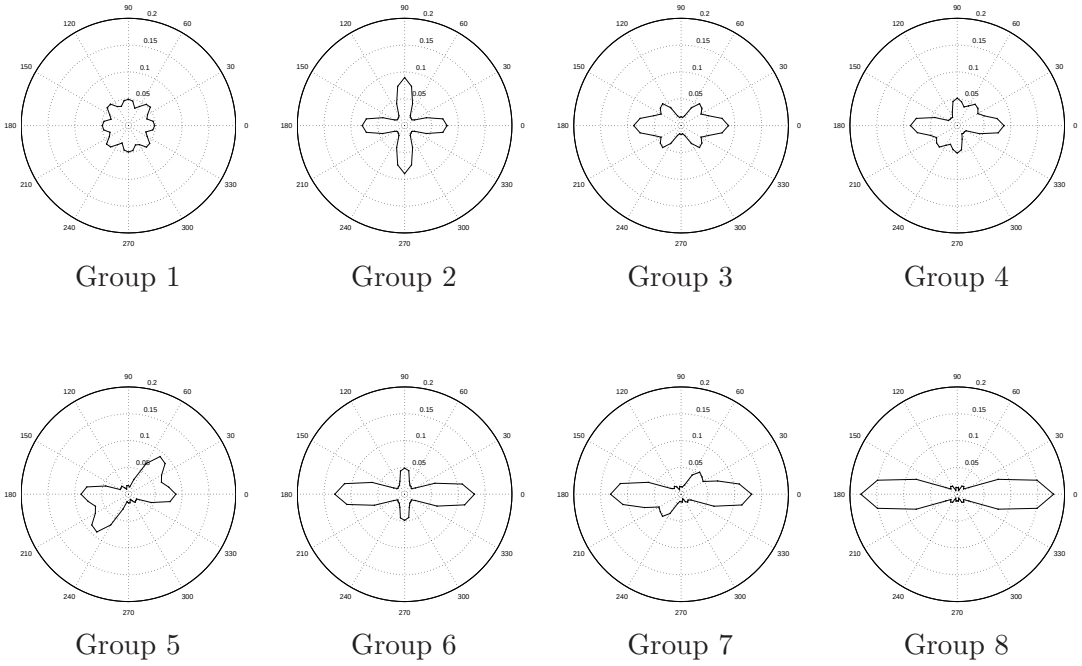


Figure C-3: Sample orientation tuning curves of S2 units in the polar coordinates: The S2 units in group 1 (cf. Table C.2) do not respond much to the bar stimuli, yielding a flat tuning curve. Group 2 shows a sharp bimodal tuning, whereas in group 5, two peaks are merged to give a larger orientation bandwidth. Groups 3 and 4 have a large node and two small nodes, while group 6 and 7 have one large node and one small node, according to the geometric configuration of the afferents. Group 8 has a sharp, unimodal tuning curve. These tuning curves represent typical results for each group.

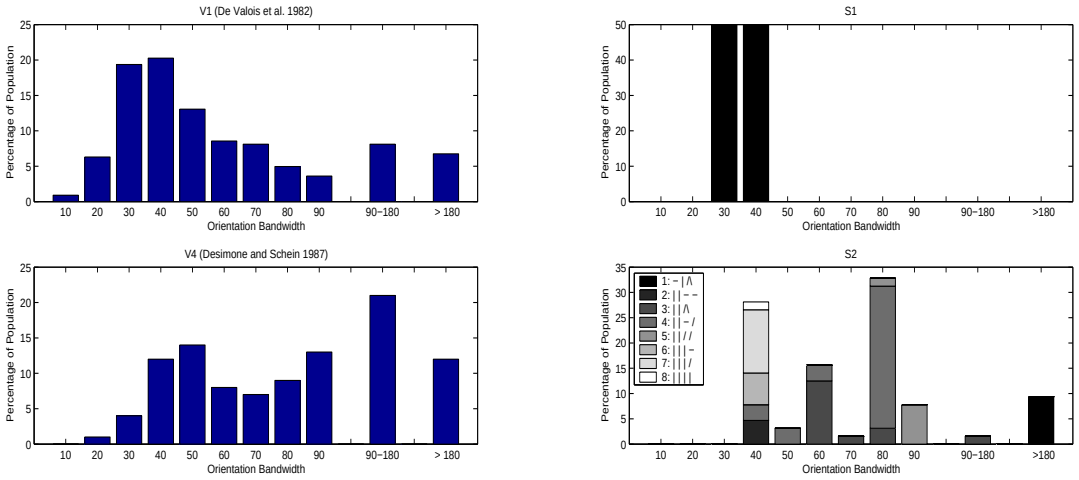


Figure C-4: Distributions of the orientation bandwidths from the physiological data (V1 and V4, taken from [De Valois et al., 1982; Desimone and Schein, 1987]) and from the simulation results (S1 and S2). The legend in the lower right histogram shows the 8-class classification scheme given in Table C.2.

to its non-parallel, orthogonal afferents. The orientation bandwidths of group 3 and 4 are quite variable because of the secondary peaks: When those secondary peaks are small, only the primary peak contributes to the orientation bandwidth. Otherwise, the secondary peaks are merged with the primary peak to yield larger orientation bandwidths. Thus, by adjusting the model parameters that influence the sharpness and the relative size of the response peaks, it is possible to obtain different bandwidth distributions. In general, the distributions are upper bounded by group 1 with the flat orientation tuning profiles and lower bounded by group 6, 7, and 8.

Fig. C-4 and Table C.1 summarize the simulation result that produced a reasonable approximation to the physiological data. Note that on average, V4 neurons and S2 units tend to have wider orientation bandwidths than V1 and S1 units. With a median bandwidth of 75°, V4 neurons have wider orientation bandwidths than V1 neurons. In the model, there is a sizable increase in the population of cells with wider bandwidths. The actual percentage values are not very close to the physiological data, since the S1 population is too simple and homogeneous (Only 11% of the S2 units in the current model are broadly tuned, whereas in V4, 33% of the neurons have wide bandwidths).

The broadening of the orientation tuning from S1 to S2 layer is observed over a wide range of model parameter values. In particular, the Gabor wave number k has a strong influence on both S1 and S2 bandwidths. Fig. C-5A shows the changing shapes of the Gabor filter at different k values. As k increases, S1 orientation bandwidth monotonically decreases. The orientation bandwidths of the S2 units also change, but rather disproportionally, as seen in Fig. C-5B. As explained before, for the S2 units in group 3 and 4, the secondary peaks in the orientation tuning profile can become significant enough and merge with the primary peaks to yield larger orientation bandwidths. When the S1 bandwidths get larger, the neighboring peaks in the S2 tuning profile are more likely to overlap, resulting in the sharp increase of the orientation bandwidths in Fig. C-5B.

Furthermore, Fig. C-5 shows that with a homogeneous population of S1 units, it is possible to consistently construct a distribution of the S2 units with wider orientation

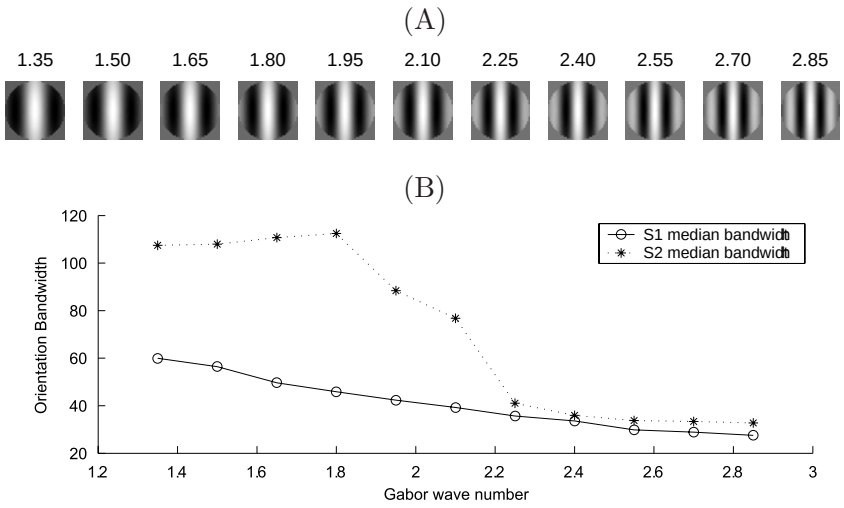


Figure C-5: (A) Gabor filters with varying wave numbers. From left to right, wave number k is increased from 1.35 to 2.85 in units of 2π. The central excitatory region becomes narrower, and the orientation bandwidth decreases, going from left to right. (B) Median orientation bandwidths of S1 and S2 units vs. the Gabor wave number k, plotted in units of 2π.

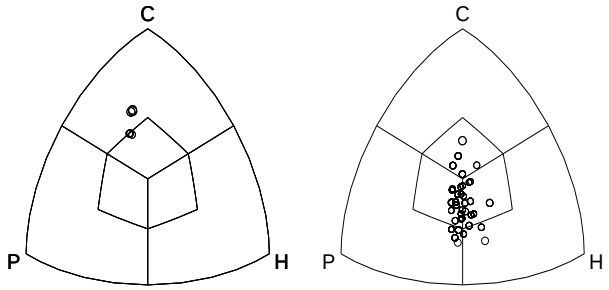
bandwidths. Thus, the increase in orientation bandwidth from V1 to V4 found in the experiments can be explained as a byproduct of cells in higher areas combining complex cell afferents.

C.4 Results: Grating Selectivity

Neurons in visual area V1 are known to be most responsive to bar-like or Cartesian stimuli, although there also appears to be a subpopulation of V1 neurons more responsive to non-Cartesian stimuli [Mahon and De Valois, 2001; Hegde and Van Essen, 2007]. In the model, the S1 population is quite homogeneous and clearly shows a bias toward Cartesian stimuli, as shown in Fig. C-6A.

Using three different classes of gratings as shown in Fig. C-2, Gallant et al. (1996) reported that the majority of neurons in visual area V4 gave comparable responses (within a factor of 2) to the most effective member of each class, while the mean responses to the polar, hyperbolic, and Cartesian gratings were 11.1, 10.0, and 8.7 spikes/second respectively, as summarized in Table C.3A. Furthermore, there was a population of neurons highly selective to non-Cartesian gratings. Out of 103 neurons, there were 20 that gave more than twice the peak responses to one stimulus class than to another: 10 showed a preference for the polar, 8 for the hyperbolic, and 2 for Cartesian gratings, as shown in Table C.3B.

When the simple S2 units (with the same set of parameters used in the orientation selectivity studies) are presented with the same set of gratings, they, as a population, exhibit a similar bias toward non-Cartesian gratings, as summarized in Table C.3. Fig. C-6B shows that there is a general trend away from the Cartesian sector, confirming the bias toward non-Cartesian stimuli. A small population of the S2 units responds significantly more to one class of stimuli than to another, as illustrated by the data points lying outside of the inner region in Fig. C-6. Note that the proportions of the cells preferring non-Cartesian



(A) S1 Responses (B) S2 Responses

Figure C-6: Responses to the three grating classes (polar, hyperbolic, and Cartesian gratings), drawn in the same convention as in Fig. 4 of [Gallant et al., 1996]. For each model unit, the maximum responses within each grating class are treated as a 3-dimensional vector, normalized and plotted in the positive orthant. This 3-dimensional plot is viewed from the $(1, 1, 1)$-direction, so that the origin corresponds to a neuron whose maximum responses to three grating classes are identical. Cartesian-preferring units will lie in the upper sector, polar in the lower left, and hyperbolic in the lower right sector. The symbols outside of the inner region correspond to the model units that gave significantly greater (by a factor of 2) responses to one stimulus class than to another. The size of each symbol reflects the maximum response obtained across the entire stimuli. Note that all S1 units (A) prefer Cartesian over polar and hyperbolic gratings, whereas most S2 units (B) lie in the lower part of the plot, indicating a general bias toward non-Cartesian gratings.

	(A)		(B)	
	Gallant (1996)	Simple S2	Gallant (1996)	Simple S2
Polar	11.1	0.14 ± 0.07	10%	10%
Hyperbolic	10.0	0.15 ± 0.06	8%	5%
Cartesian	8.7	0.05 ± 0.04	2%	0%

Table C.3: (A) Mean responses to three different classes of gratings. Physiological data are in units of spikes/second, whereas the model responses (baseline-subtracted) lie between 0 and 1. Although a direct comparison of the numerical values is meaningless, the model units and the V4 neurons both show a clear bias toward non-Cartesian gratings. (B) Percentage of cells that gave more than twice the peak responses to one stimulus class than to another.

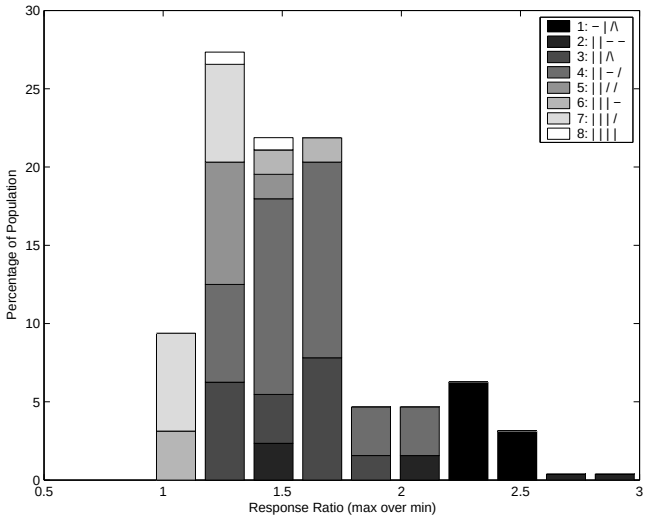


Figure C-7: Distribution of the response ratio (maximum over minimum) to three grating classes. The ratio of 1 indicates that the cell gave the same maximum responses to all three grating classes.

gratings in the model and in the experiment agree quite well, in this simplified setup.

The ratio of the maximum and the minimum responses to three grating classes shows that most S2 units (82%, very close to the estimate of 80% in [Gallant et al., 1996]) respond to all three types of gratings comparably (within a factor of two) as seen in Fig. C-7. However, for a small fraction of cells, this maximum-over-minimum ratio exceeds 2, indicating an enhanced selectivity toward one class of stimuli. In particular, the S2 units in group 1 (Table C.2) stand out in the distribution, since they respond very weakly to the Cartesian stimuli, but strongly to the non-Cartesian stimuli.

Fig. C-8 shows the distribution of the S2 unit responses, along with the 8-class classification scheme (Table C.2). They illustrate that the S2 units in group 8, whose afferents are pointing in parallel orientations, produce large responses to Cartesian gratings, as expected. On the other end of the spectrum, the S2 units in group 1, whose pooled afferents are selective to different orientations, show higher responses to non-Cartesian gratings.

The average response of the population to each grating is plotted in Fig. C-9A, where the bias in favor of non-Cartesian stimuli is again apparent. In a good qualitative agreement with Figure 3D of [Gallant et al., 1996], the average population responses are high for polar and hyperbolic gratings of low/intermediate frequencies. Within the Cartesian stimulus space, the average response is also peaked around the low/intermediate frequency region. The concentric grating of low frequency (marked with *) shows the maximum average response. For reference, Fig. C-9B through C-9D show the tuning curves of three individual S2 units that are most selective to each grating type.

One of the major differences between the physiological data in [Gallant et al., 1996] and the aforementioned simulation results is the lack of highly selective S2 units to one stimulus class only (In the scatter plot, those units would lie along the direction of $(1, 0, 0)$, $(0, 1, 0)$, or $(0, 0, 1)$). In fact, as seen in Fig. C-6B, most of the S2 units lie near the boundary between the polar and hyperbolic sectors, meaning that they respond quite similarly to these gratings, but differently to Cartesian gratings.

The above result therefore suggests that the 2x2 arrangement of the Gabor-like fea-

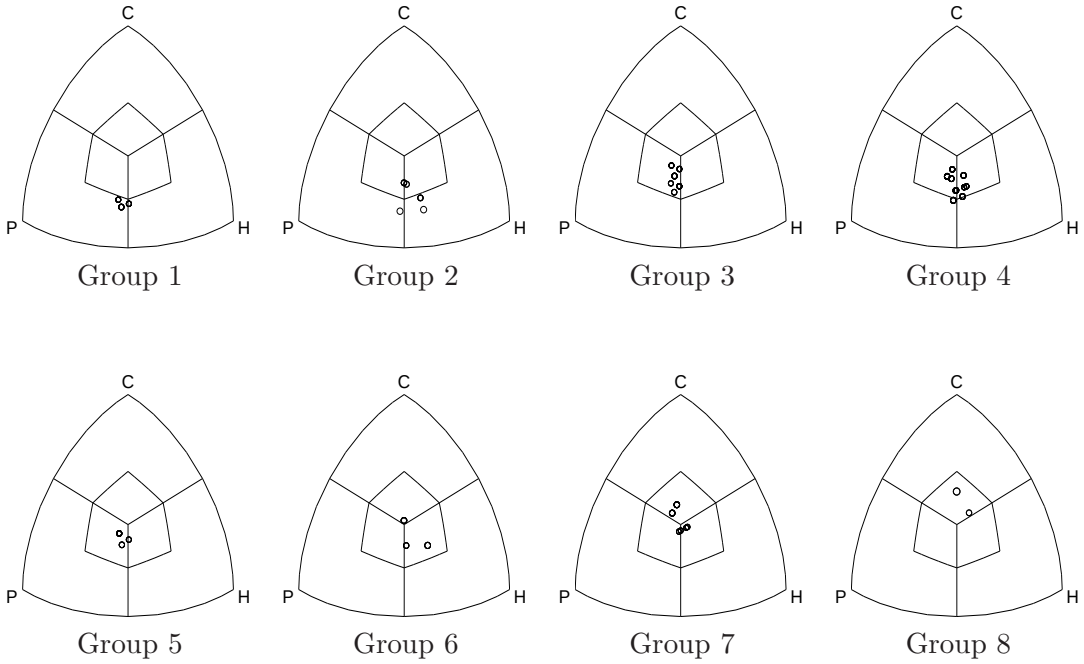


Figure C-8: When all 256 simple S2 units are plotted in the same format as Fig. C-6, it is apparent that group 1 and 2 are composed of highly non-Cartesian units, while the preference for Cartesian stimuli slowly increases toward group 8.

tures may be too simplistic, possibly because the sampling of the adjacent afferents is too correlated to contain enough distinguishing features across the polar-hyperbolic dimension. The construction of the S2 units can be extended to investigate these issues. The feature complexity can be increased by using different combination schemes (e.g., 3x3) or by sampling the afferents from non-adjacent regions. Fig. C-10 illustrates that by introducing such modifications, it is possible to obtain more uniformly distributed responses in the polar-hyperbolic-Cartesian space, while maintaining a general bias toward non-Cartesian stimuli. This result indicates that combining non-local, less-correlated features would be important in building features that can distinguish object classes better (in this case, polar vs. hyperbolic gratings).

Using more C1 afferents, it is also possible to introduces other variants of S2 units with different grating selectivities. Using a 3x3 grid for feature combination with 4 different orientations yields $4^9 = 262144$ possibilities. However, by increasing the number of the afferents, the bias toward non-Cartesian grating is also increased, since it is less likely to have most of the afferents with the same orientation selectivities.

C.5 Discussion

In this section, a population of simple S2 units was used to provide some intuitive illustration of how the observed progress in shape selectivity from V1 to V4 may arise. Physiological experiments show that along the ventral pathway, the selectivity for non-Cartesian stimuli increases. It has been reported that there are more neurons responsive to non-Cartesian

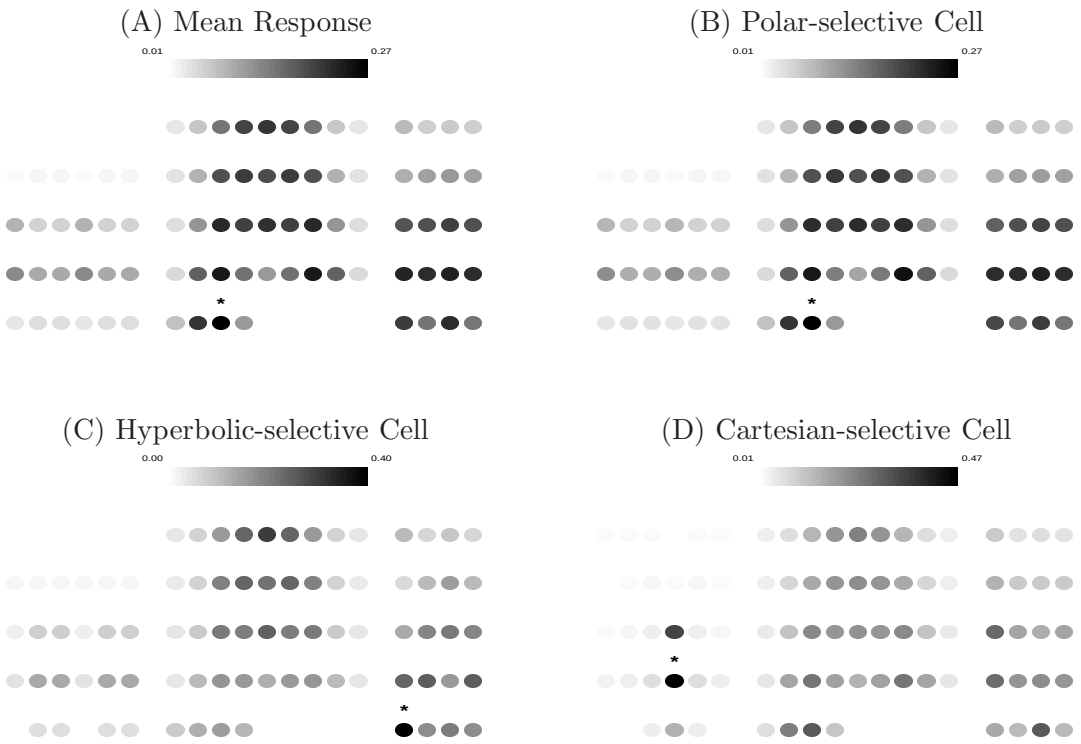
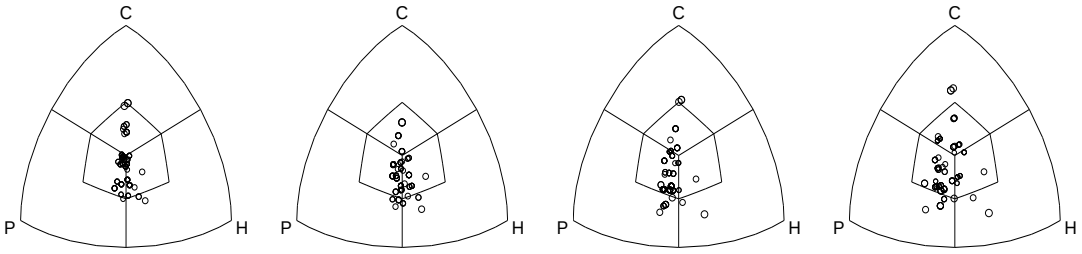


Figure C-9: Here, average population responses (A) and three sample tuning curves most selective to each of the three grating classes (B, C, D) are shown. The responses are arranged in the same layout as in Fig. C-2. The most effective stimulus is marked with an asterisk (*) on top.

(A) 2x2 Scheme



(B) 3x3 Scheme

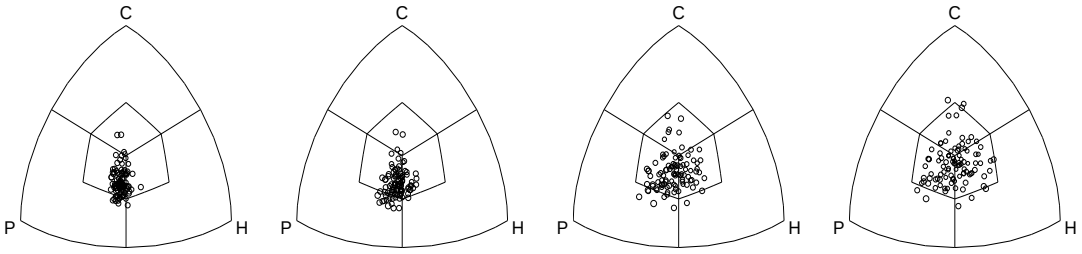


Figure C-10: Population responses to three grating classes, from 100 S2 units that are chosen randomly from all possible feature combinations. Top row (A) shows the results with the 2x2 feature combination scheme, and the bottom row (B) shows the 3x3 scheme. Going from left to right, the distance between the C1 afferents is increased. In the first column, the C1 afferents are partially (1/2) overlapping. In the second column, the C1 afferents are adjacent. Thus, the plot in the top row of the second column represents the result using the standard HMAX parameters. In the third and the fourth columns, the C1 afferents are even farther apart (1 or 2 times the C1 pooling range). As the distance between the C1 afferents increases, the features combined in one S2 receptive field are sampled from farther regions of the stimulus image.

gratings in V2 and V4 than in V1 [Mahon and De Valois, 2001; Hegde and Van Essen, 2007]. Furthermore, the neurons in the visual area V4 show enhanced selectivity for non-Cartesian gratings, and very few neurons are highly responsive to Cartesian gratings only [Gallant et al., 1996]. A similar trend is apparent even in the simple population of S2 units, or rather it has been implicitly built into it, by combining oriented filters (naturally responsive to Cartesian gratings) into non-parallel, non-Cartesian features. In the S1 layer, there is no model unit more responsive to non-Cartesian gratings, whereas in the S2 layer the majority prefers non-Cartesian gratings.

The model posits that the increase in complexity results from simple combinations of complex cell afferents. Despite its simplicity, the population of the simple S2 units came quite close to several physiological data in V4. In particular, the model exhibited the broadening of the orientation bandwidth and the bias toward non-Cartesian stimuli, while successfully reproducing some of the population statistics. Even a simple 2x2 combination of the afferents could yield fairly complex behaviors in the population. Furthermore, it was noted that the model units whose afferents were non-parallel and orthogonal served to yield wide orientation bandwidth and high selectivity for non-Cartesian stimuli.

The grating stimuli revealed some discrepancies between the simplified S2 units and the physiological data, as there was a lack of model units strongly selective for either polar or hyperbolic gratings only. More complex S2 units were obtained by increasing the spatial separation of the C1 afferents. This provides an interesting prediction for experiments regarding the receptive field substructure of the neurons in higher visual areas, for which there are some preliminary experimental evidences in V2 [Anzai et al., 2002]. Interestingly, features based on spatially-separated, complex cell-like afferents have been previously postulated based on computational grounds [Amit and Geman, 1997]. An alternative, more trivial way to obtain cells strongly selective for non-Cartesian gratings, even though not explored here, would be to assume more complex, non-Cartesian S1 features that are more selective toward the features found in the stimuli set. Physiological results indicate that V1 indeed contains the neurons responsive to radial, concentric, or hyperbolic gratings [Mahon and De Valois, 2001; Hegde and Van Essen, 2007]. Finally, it appears that the bar and grating stimuli are too limited as a stimulus set to provide strong constraints for the model, as even the simplified S2 units seemed to have enough degrees of freedom to cover various bandwidth distributions and grating selectivities.

Appendix D

More on Clutter Effects and Tradeoff

D.1 Clutter Effects with Unknown Center of Tuning

Unlike the situation presented in Fig. 5-6, the Gaussian-like tuning operation may not be centered exactly at one of the objects in the stimulus set. However, if the optimal input pattern for a model unit is very different from the input patterns generated by the experimental stimuli, the neural response is likely to be too small to be recorded in the experiment, as shown in the left panel in Fig. D-1B. An opposite case is shown in the right panel of Fig. D-1B, where a high response is measured, regardless of the stimuli (i.e., the model unit shows no selectivity).

Occasionally, a model unit may produce non-trivial responses to the experimental stimuli (i.e., it shows some selectivity), even when it is actually tuned to some object not contained in the experimental stimulus set. In such cases (like the middle panel in Fig. D-1B), the suppressive clutter effects are still observed, especially when the most effective and the least effective stimuli are paired. Sometime, the paired stimuli may enhance the responses and yield a facilitatory effect. The overall direction of the effects (facilitatory vs. suppressive) depends on where the true center of tuning is and how the afferent activity changes with respect to this center. If the afferent activity to the clutter condition moves closer to the optimal pattern of afferent activation, the clutter effect will be facilitatory.

Relatively large ranges of selectivity and tolerance properties and their tradeoff behaviors can also be observed from a population of model units tuned to randomly chosen afferent activation patterns, even when all the other parameters are fixed (e.g., same numbers of afferents, same tuning width, etc.), as shown in Fig. D-2. However, in the experiment of [Zoccolan et al., 2007], most recorded neurons produced sizable responses to at least one of the stimuli, indicating that they were likely to be tuned rather close to those objects, and the facilitatory clutter effects were rare (unlike Fig. D-1B). Furthermore, in the physiological data, there was no correlation between the selectivity (sparseness index) and the maximum response of the neural population, in contrast with the simulation results in Fig. D-2. Therefore, the tradeoff result from physiological data is not likely to be due solely to the random selectivity of the recorded IT neurons.

Possibly, the IT neurons may acquire explicit representations for the frequently encountered stimuli. The clutter condition typically combines random and uncorrelated pairs of the single objects, and, hence, the newly acquired representation in the cortex is likely to

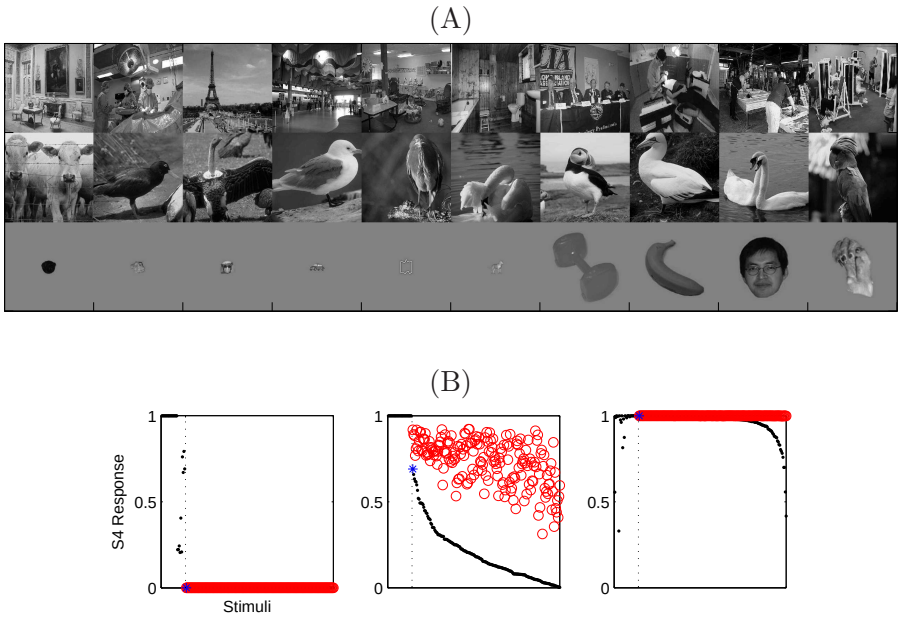


Figure D-1: Instead of recording from the model units whose optimal stimuli belong to the experimental stimulus set, a different population is tested on the stimulus set. (A) shows the examples of the target stimuli: pictures of scenes, animals, and isolated objects at different scales. These images are taken from the databases of [Serre et al., 2005, 2007a; Hung et al., 2005]. (B) shows the response of three different model units that are selective for the 6th, 17th, and 26th images in (A), respectively. The responses to the 30 images in (A) are shown on the left side of the dotted line, and those to the main stimulus set (Fig. 5-3A) are shown on the right, and the clutter responses are shown in red dots. Because the response to the stimulus set was too small with the usual model parameters, the sigmoid parameters were adjusted to make these units more broadly tuned ($\beta = 0.5$ instead of the usual 0.9, in Eq. 2.9). Hence, these model units respond strongly to most of the images in (A) and show rather trivial response profiles as shown in the left and right panels in (B). One of the rare, nontrivial examples is presented in the middle panel, where a variety of clutter effects, both suppressive and facilitatory, are observed. Most clutter effects, especially in combination with the worse stimuli, are suppressive (i.e., most red dots are below the *), as observed in the experiment [Zoccolan et al., 2007].

be based on the individual object. However, if the clutter condition is learned (i.e., the neurons are tuned to the paired stimuli presented together), the clutter effect would be more facilitatory than suppressive. The caveat of this prediction is that (1) it will be difficult to find a neuron with such a designated preference for the clutter condition, even after a long training period (for example, in the experiment of [Logothetis et al., 1995], a very small subpopulation of recorded neurons showed tuning to the trained objects) and (2) the learning rule in cortex may not be a simple "imprinting" or memory of the frequently encountered stimuli. Instead, the visual cortex may employ more sophisticated strategies to acquire more abstract, component-based representations from the visual experience (for example, in the experiment of [Zoccolan et al., 2005], most neurons were selective for the extreme examples or "caricatures" within the parametrized stimulus set).

D.2 Factors Affecting the Tradeoff

There are multiple parameters that determine the response properties (like the sharpness of tuning) of a model unit. In this section, a few such parameters that may be more biologically relevant are listed, and it is shown that the tradeoff between selectivity and invariance is robustly observed under the variations of these parameters.

1. Increasing the number of afferent units (the dimensionality of the input) can sharpen the neural tuning, as shown in Fig. 5-4. An intuitive argument is that when there are more afferents, the total change in the input due to the stimulus change is likely to be higher and the corresponding output will be perturbed by a larger amount. An alternative explanation is that a given tuning width is effectively smaller in a higher dimensional space, because a distance measure scales with the dimensionality (e.g., the Euclidean distance scales as the square root of the number of dimension). Furthermore, a Gaussian-like tuning operation represents the conjunctions of multiple features, so that, with many inputs, it is more difficult to produce the suitable combination of features and generate a large response. Hence, the model unit with many afferents will be unresponsive to many stimuli, and its response profile will be rather sparse and highly selective.

2. A model IT unit performs a Gaussian-like tuning operation over a set of afferent inputs, and its final output may be subject to a sigmoid monotonic nonlinearity, Eq. 2.9. Similar to the role of σ in a Gaussian function $y = \exp^{(-|x-w|^2/2\sigma^2)}$, the sigmoid parameter α and β can control the width of the tuning curve, or the threshold and the steepness of nonlinearities in the neural circuit implementing the tuning operation, as shown in Fig. D-3A.

3. The dynamic range of the tuning function probed by the experimental stimuli also affects the sensitivity profile of the neural response. For example, the local slope (first derivative) of the bell-shaped Gaussian function is small at the points near or far away from its tuning center, but large in the intermediate ranges. A sigmoid-like nonlinearity also yields the largest variation per unit change in the input near the intermediate ranges (near β in Eq. 2.9). Therefore, for a given stimulus set, depending on the range at which the experimental stimuli are probing the tuning curve, the neural response may produce different ranges of selectivity and invariance behaviors. However, if the difference between the maximum and the minimum responses is fairly

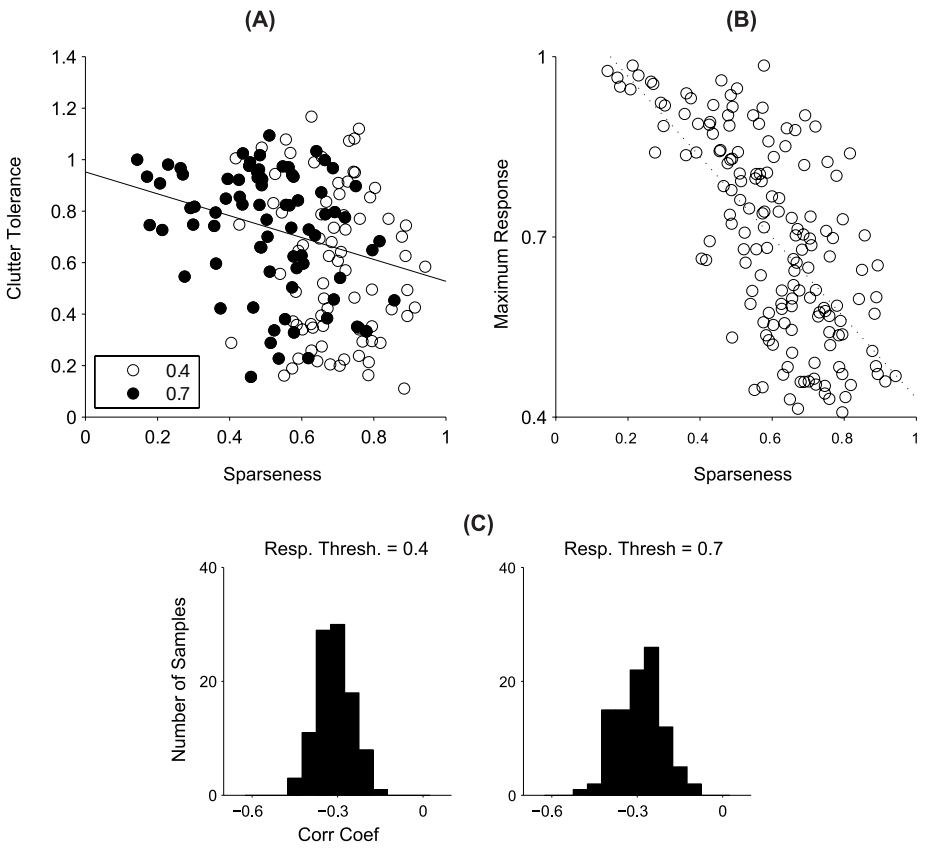


Figure D-2: (A) 300 random tuning centers are selected within a 50-dimensional input space (i.e., 50 afferent inputs for each model IT unit, which is less than the usual number of inputs in other simulations). Reducing the number of afferents and using less steep sigmoid function (in Eq. 2.9, (α, β) was set to (30, 0.6), instead of the usual value of (40, 0.9)) helped to ensure that the model responses were not too small and that a wide range of selectivity and tolerance behaviors was covered. Only a subset of model units that had the maximum responses higher than 0.4 (all circles) or 0.7 (filled circles only) to the experimental stimulus set was included in the analysis. For both subsets of model units, tradeoff between selectivity and clutter tolerance was observed, although much noisier (r = -0.32 and -0.29, for the populations fulfilling the 0.4 and 0.7 response threshold, respectively) than observed for the recorded IT neuronal population and other model simulation results. (B) Unlike the results from the recorded IT neurons [Zoccolan et al., 2007] and other simulation results (Fig. 5-4D and Fig. D-3), there was a strong correlation (r = -0.68) between the sparseness index and the maximum response of the model units. (C) To better quantify at which extent the random tuning centers could account for the tradeoff observed in the neural data, the sampling of the random tuning centers was repeated 100 times to obtain 100 sets of 300 model units, and the distribution of the resulting correlation coefficients between sparseness and clutter tolerance was computed. The correlation between sparseness and clutter tolerance (mean = -0.3, for both threshold values) was significantly weaker (p = 0.01) than the correlation computed for the IT neuronal population (r = -0.46). Overall, these analyses suggest that the range of selectivity and clutter tolerance measured for the recorded neuronal population and the tradeoff between these properties cannot be accounted by the model if only the tuning centers of the model IT units are varied.

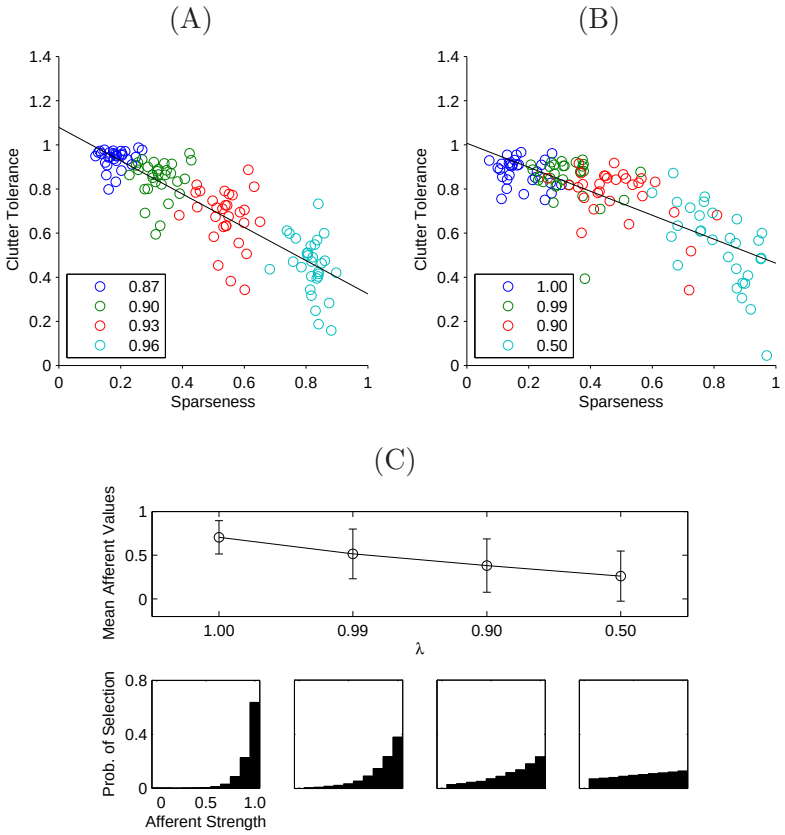


Figure D-3: These figures show the tradeoff between selectivity and clutter tolerance of the S4 units in the model, as the number of afferent units is fixed at 500 and other parameters are varied. The sigmoid parameter β in Eq. 2.9 is varied in (A), and the selection rule on the afferent neurons is varied in (B). The correlation coefficients between the sparseness (selectivity) and the tolerance indices were -0.88 and -0.79, respectively. As illustrated in (C), the afferent selection rule is systematically varied by parameterizing the probability of the afferent selection bias (with an exponential probability distribution with parameter λ) according to the afferent responses to a target stimulus. The exponential distributions are generated by $-\log(1-p\lambda)$ where p is a uniform distribution on the afferent neurons sorted according to the responses to the target stimulus. Therefore, as shown in the bottom four panels, the probability distribution becomes more biased with increasing λ, and the mean response to the target increases. In other words, with larger values of λ, the afferent neurons more responsive to the target stimuli are selected, and the clutter tolerance increases, while the selectivity decreases, as shown in (B).

large (as in the case of the experimental data [Zoccolan et al., 2007]), the stimulus set is likely to be rich enough, and is probing the full range of the tuning function and reporting the true selectivity of the neuron.

4. The center of tuning can affect the selectivity of the neural response. As shown by the spread of points in Fig. 5-4 and Fig. D-2, a model unit tuned to a particular stimulus may have different selectivity or sparseness index from another unit tuned to a different stimulus, even when all the other parameters (like the number of afferents or the sigmoid parameters) are the same. If a neuron is tuned to a simple and common feature (like an oriented edge which is found in most images), it would be less sensitive to the changes in the stimulus. Similarly, if a neuron is tuned to more complex features, it will be more sensitive. The gradual progression from simple to complex shape selectivities and from small to large ranges of invariance is likely to be varied and overlapped within the hierarchy of the visual cortex. In other words, within each layer of the hierarchy, there is a continuous spectrum of selectivity and invariance, overlapping with those of other neighboring areas [Mahon and De Valois, 2001; Hegde and Van Essen, 2007]. As a result, the population of neurons from the same area will exhibit an inherently wide range of tuning behaviors. In the real physiological experiments, however, the technical time constraints inevitably limits the number of stimuli and such a diversity of tuning is difficult to probe in detail. Also note that, as shown in Fig. D-2, the diversity of the tuning centers alone may not account for the full tradeoff results in [Zoccolan et al., 2007].

5. Another factor that influences the selectivity and clutter tolerance behavior is the selectivity and tolerance properties of the afferent neurons themselves. If they are already highly sensitive to the input changes, the efferent neuron receiving those responses will correspondingly be more sensitive to the cluttered stimuli. Therefore, the afferent selection rule that determines the properties of the afferent neurons would play a role in shaping the response profile of the efferent neuron. For example, suppose that a neuron makes synapses with the afferent neurons that produce large responses to the stimulus set, especially to the target stimulus. These afferent neurons would tend to produce higher responses to the most stimuli in the set (i.e., their selectivities are low). As a result, the efferent neuron itself will also be less sensitive and more tolerant to the stimulus changes. As shown in Fig. D-3B and D-3C, it is possible to get a different range of tradeoff behavior with different degrees of selection bias on the afferent neurons. Note that the simulation result in Fig. 5-4D is also related to the result in Fig. D-3B. When the top most activated afferents are chosen, increasing the number of afferent neurons tends to add the ones that are more susceptible to the clutter or the stimulus change, while reducing the mean afferent responses, similar to Fig. D-3C.

As discussed above, multiple, sometimes related, factors contribute to the clutter effects, even in a relatively simple model that only assumes a feedforward hierarchy of the Gaussian-like and max-like operations. In particular, the tradeoff is a robustly observed phenomena (Fig. 5-4 and D-3), even when different parameters and factors affect the selectivity and invariance properties of each neuron.

D.3 Dependence on the Receptive Field Size

The clutter effect is influenced by the size of receptive field with respect to the stimulus sizes and to the distances between the stimuli in the clutter condition. An inferior temporal neuron typically has a large receptive field in which multiple objects can appear. The receptive fields of the neurons in earlier cortical areas tend to be smaller, and, hence, cover smaller portions of a visual scene.

In the model, the IT-like model units and their afferent units (i.e., S4 and C2b units) are assumed to have large receptive fields, but the afferents to the afferent neurons (S2b units) cover a smaller portion of the visual scene. In the case where the receptive fields of the S2b units are small and confined to a single object (left side in Fig. D-4A), their individual responses to the clutter will be equal to the responses to the single objects, as there would not be much interference by the cluttering object within the receptive field. On the other hand, in the case where the S2b units cover larger visual angles (right side in Fig. D-4A), multiple objects can co-appear within the receptive fields and perturb the neural responses more, resulting in clutter intolerance. Such dependency is shown in Fig. D-4B. Therefore, it is expected that as the distance between the cluttering objects increases, the clutter tolerance would also increase, as long as the increased inter-stimulus distance is large enough with respect to the receptive field sizes of the afferent neurons.

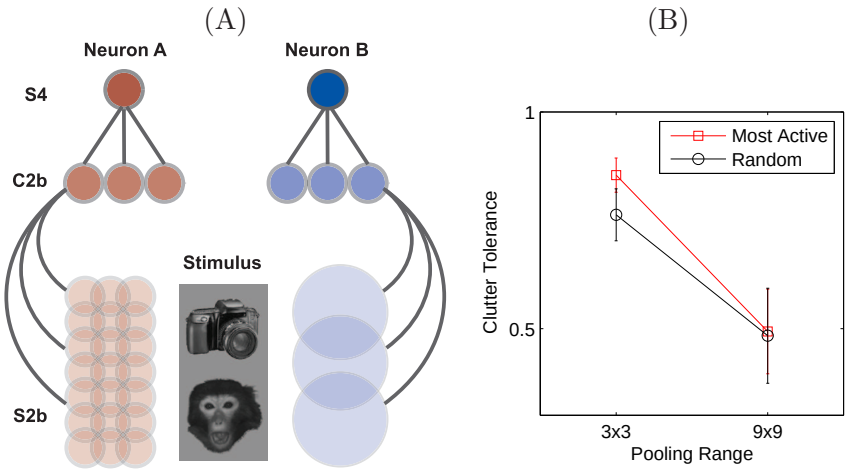


Figure D-4: (A) Suppose that the receptive field of an efferent neuron is built up by pooling from the afferent neurons with different receptive field sizes. The smaller receptive fields (left) typically cover smaller portions of a visual scene within a single object, and the larger receptive fields (right) may cover multiple objects. In the latter case, the clutter condition will generally produce greater perturbations in the neural responses, so the clutter tolerance will be smaller. In the model, such different situations can be modeled by two different types of S2b units. The first type (with smaller receptive fields) pools over ten C1 units within a 3x3 spatial grid, and the other type (with larger receptive fields) pools over the same number of C1 units within a 9x9 spatial grid (note that in other typical simulations, 100 C1 units within a 9x9 grid are pooled). In terms of visual angle, the first type covers between 0.7 and 2.8 degrees, and the second, between 1.4 and 5.4. Each stimulus spanned about 2 degrees of visual angle, and in the clutter condition, two stimuli were separated by 2.5 degrees (center to center distance). These two different types of S2b units are pooled by a soft-max operation into different C2b and then S4 units, whose receptive fields cover the entire visual field. The S4 units, corresponding to Neuron A or B in the panel (A), receive inputs from 500 C2b units. As shown in (B), the clutter tolerance decreases when the C2b units receive the inputs from the afferent neurons with larger receptive fields (i.e., the second type of S2b units with a 9x9 pooling grid). The afferent selection rule (selecting random vs. most active afferent units) has a stronger influence on the clutter tolerance in the case with smaller receptive fields.

Bibliography

E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America*, 2(2):284–299, 1985.

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1588, 1997.

A. Anzai, D. C. Van Essen, X. Peng, and J. Hegde. Receptive field structure of monkey V2 neurons for encoding orientation contrast. *Journal of Vision*, 2(7), 2002.

M. S. Bartlett and T. J. Sejnowski. Learning viewpoint-invariant face representations from visual experience in an attractor network. *Network*, 9(3):399–417, 1998.

K. T. Blackwell, T. P. Vogl, and D. L. Alkon. Pattern matching in a model of dendritic spines. *Network*, 9(1):107–21, 1998.

L. J. Borg-Graham, C. Monier, and Y. Fregnac. Visual input evokes transient and strong shunting inhibition in visual cortical neurons. *Nature*, 393(6683):369–73, 1998.

A. Borst, M. Egelhaaf, and J. Haag. Mechanisms of dendritic integration underlying gain control in fly motion-sensitive interneurons. *J Comput Neurosci*, 2(1):5–18, 1995.

S. L. Brincat and C. E. Connor. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience*, 7:880–886, 2004.

S. L. Brincat and C. E. Connor. Dynamic shape synthesis in posterior inferotemporal cortex. *Neuron*, 49(1):17–24, 2006.

R. M. Bruno and D. J. Simons. Feedforward mechanisms of excitatory and inhibitory cortical receptive fields. *Journal of Neuroscience*, 22(24):10966–75, 2002.

C. Cadieu, M. Kouh, A. Pasupathy, C. E. Connor, M. Riesenhuber, and T. Poggio. A model of V4 shape selectivity and invariance. *In Submission*, 2007.

M. Carandini and D. J. Heeger. Summation and division by neurons in primate visual cortex. *Science*, 264(5163):1333–6, 1994.

M. Carandini, D. J. Heeger, and J. A. Movshon. Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, 17(21):8621–8644, 1997.

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

F. S. Chance and L. F. Abbott. Divisive inhibition in recurrent networks. *Network*, 11(2): 119–29, 2000.

F. S. Chance, S. B. Nelson, and L. F. Abbott. Complex cells as cortically amplified simple cells. *Nature Neuroscience*, 2(3):277–82, 1999.

H. J. Chisum and D. Fitzpatrick. The contribution of vertical and horizontal connections to the receptive field center and surround in V1. *Neural Netw.*, 17:681–693, 2004.

Y. Dan and M. M. Poo. Spike timing-dependent plasticity of neural circuits. *Neuron*, 44 (1):23–30, 2004.

J. G. Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20(10):847–856, 1980.

S. V. David, B. Y. Hayden, and J. Gallant. Spectral receptive field properties explain shape selectivity in area V4. *Journal of Neurophysiology*, 96(6):3492–3505, 2006.

P. Dayan and L. Abbott. *Theoretical Neuroscience*. MIT Press, 2001.

R. L. De Valois, E. W. Yund, and N. Hepler. The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, 22(5):531–44, 1982.

P. De Weerd, R. Desimone, and L. G. Ungerleider. Cue-dependent deficits in grating orientation discrimination after V4 lesions in macaques. *Visual Neuroscience*, 13(3):529–38, 1996.

R. Desimone and S. Schein. Visual properties of neurons in area V4 of the macaque: Sensitivity to stimulus form. *Journal of Neurophysiology*, 57:835–868, 1987.

R. J. Douglas, K. A. Martin, and D. Whitteridge. A canonical microcircuit for neocortex. *Neural Computation*, 1:480–488, 1989.

D. Felleman and D. C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex*, 1:1–47, 1991.

D. Ferster and K. D. Miller. Neural mechanisms of orientation selectivity in the visual cortex. *Ann. Rev. Neurosci*, 23:441–71, 2000.

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

L. M. Frank, E. N. Brown, and M. A. Wilson. A comparison of the firing properties of putative excitatory and inhibitory neurons from CA1 and the entorhinal cortex. *Journal of Neurophysiology*, 86(4):2029–40, 2001.

D. J. Freedman and J. A. Assad. Experience-dependent representation of visual categories in parietal cortex. *Nature*, 443(7107):85–8, 2006.

D. J. Freedman, M. Riesenhuber, T. Poggio, and E. K. Miller. A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience*, 23(12):5235–46, 2003.

W. A. Freiwald, D. Y. Tsao, R. B. Tootell, and M. S. Livingstone. Single-unit recording in an fMRI-identified macaque face patch. II. coding along multiple feature axes. In *Society for Neuroscience*, volume Program No. 362.6, Washington, DC, 2005.

W. A. Freiwald, D. Y. Tsao, R. B. H. Tootell, and M. S. Livingstone. Complex and dynamic receptive field structure in macaque cortical area V4d. *Journal of Vision*, 4(8):184–184, 2004.

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

J. L. Gallant, C. E. Connor, S. Rakshit, J. W. Lewis, and D. C. Van Essen. Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *Journal of Neurophysiology*, 76(4):2718–39, 1996.

J. L. Gallant, R. E. Shoup, and J. A. Mazer. A human extrastriate area functionally homologous to macaque V4. *Neuron*, 27(2):227–35, 2000.

T. J. Gawne and J. M. Martin. Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. *Journal of Neurophysiology*, 88(3):1128–35, 2002.

P. Girard, S. G. Lomber, and J. Bullier. Shape discrimination deficits during reversible deactivation of area V4 in the macaque monkey. *Cereb. Cortex*, 12(11):1146–56, 2002.

F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269, 1995.

C. G. Gross, C. E. Rocha-Miranda, and D. B. Bender. Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of Neurophysiology*, 35(1):96–111, 1972.

S. Grossberg. Contour enhancement, short-term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, 52(3):213–257, 1973.

R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–51, 2000.

D. J. Heeger. Modeling simple-cell direction selectivity with normalized, half-squared, linear operators. *Journal of Neurophysiology*, 70(5):1885–98, 1993.

J. Hegde and D. C. Van Essen. Strategies of shape representation in macaque visual area V2. *Visual Neuroscience*, 20(3):313–28, 2003.

J. Hegde and D. C. Van Essen. Temporal dynamics of shape analysis in macaque visual area V2. *Journal of Neurophysiology*, 92(5):3030–42, 2004.

J. Hegde and D. C. Van Essen. A comparative study of shape representation in macaque visual areas v2 and v4. *Cereb. Cortex*, 17(5):1100–16, 2007.

D. Hinkle and C. E. Connor. Three-dimensional orientation tuning in macaque area V4. *Nature Neuroscience*, 5:665–670, 2002.

G. R. Holt and C. Koch. Shunting inhibition does not have a divisive effect on firing rates. *Neural Computation*, 9(5):1001–1013, 1997.

D. Hubel and T. Wiesel. Receptive fields and function architecture of monkey striate cortex. *Journal of Physiology*, 195:215–243, 1968.

D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–54, 1962.

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

C. P. Hung, G. Kreiman, T. Poggio, and J. J. DiCarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–6, 2005.

M. Ito, H. Tamura, I. Fujita, and K. Tanaka. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, 73(1):218–26, 1995.

J. P. Jones and L. A. Palmer. The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1187–211, 1987.

U. Knoblich, J. Bouvrie, and T. Poggio. Biophysical models of neural computation: Max and tuning circuits. Technical Report CBCL Paper, Massachusetts Institute of Technology, 2007.

E. Kobatake and K. Tanaka. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, 71(3):856–67, 1994.

M. Kouh and T. Poggio. A general mechanism for cortical tuning: Normalization and synapses can create gaussian-like tuning. Technical Report CBCL Paper #245/AI Memo #2004-031, Massachusetts Institute of Technology, 2004.

M. Kouh and T. Poggio. A canonical cortical circuit for nonlinear operations. *In Submission*, 2007.

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

G. Kreiman. Neural coding: computational and biophysical perspectives. *Physics of Life Reviews*, 1(2):71–102, 2004.

I. Lampl, D. Ferster, T. Poggio, and M. Riesenhuber. Intracellular measurements of spatial integration and the max operation in complex cells of the cat primary visual cortex. *Journal of Neurophysiology*, 92(5):2704–13, 2004.

D. K. Lee, L. Itti, C. Koch, and J. Braun. Attention activates winner-take-all competition among visual filters. *Nature Neuroscience*, 2(4):375–81, 1999.

T. S. Lee and D. Mumford. Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America*, 20(7):1434–48, 2003.

D. A. Leopold, I. V. Bondar, and M. A. Giese. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, 442(7102):572–5, 2006.

F. F. Li, R. Van Rullen, C. Koch, and P. Perona. Rapid natural scene categorization in the near absence of attention. *Proc. Natl. Acad. Sci. USA*, 99(14):9596–601, 2002.

Z. Li. A neural model of contour integration in the primary visual cortex. *Neural Computation*, 10(4):903–40, 1998.

M. S. Livingstone and B. R. Conway. Substructure of direction-selective receptive fields in macaque V1. *Journal of Neurophysiology*, 89(5):2743–59, 2003.

N. K. Logothetis, J. Pauls, and T. Poggio. Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5):552–63, 1995.

L. Mahon and R. De Valois. Cartesian and non-Cartesian responses in LGN, V1, and V2 cells. *Visual Neuroscience*, 18:973–981, 2001.

N. J. Majaj, M. Carandini, and J. A. Movshon. Motion integration by neurons in macaque MT is local, not global. *Journal of Neuroscience*, 27(2):366–70, 2007.

J. Marino, J. Schummers, D. C. Lyon, L. Schwabe, O. Beck, P. Wiesing, K. Obermayer, and M. Sur. Invariant computations in local cortical networks with balanced excitation and inhibition. *Nature Neuroscience*, 8(2):194–201, 2005.

M. Maruyama, F. Girosi, and T. Poggio. A connection between GRBF and MLP. Technical Report AI Memo 1291, Massachusetts Institute of Technology, 1992.

J. A. Mazer and J. L. Gallant. Goal-related activity in V4 during free viewing visual search. Evidence for a ventral stream visual salience map. *Neuron*, 40(6):1241–50, 2003.

F. Mechler and D. L. Ringach. On the classification of simple and complex cells. *Vision Research*, 42(8):1017–33, 2002.

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

B. W. Mel and J. Fiser. Minimizing binding errors using learned conjunctive features. *Neural Computation*, 12(2):247–278, 2000.

B. W. Mel, D. L. Ruderman, and K. A. Archie. Translation-invariant orientation tuning in visual complex cells could derive from intradendritic computations. *Journal of Neuroscience*, 18(11):4325–34, 1998.

W. H. Merigan and H. A. Pham. V4 lesions in macaques affect both single- and multiple-viewpoint shape discriminations. *Visual Neuroscience*, 15(2):359–67, 1998.

K. D. Miller. Synaptic economics: competition and cooperation in synaptic plasticity. *Neuron*, 17(3):371–4, 1996.

J. Moody and C. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):289–303, 1989.

V. B. Mountcastle. Introduction: Computation in cortical columns. *Cereb. Cortex*, 13(1): 2–4, 2003.

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

J. Mutch and D. G. Lowe. Multiclass object recognition with sparse, localized features. In *IEEE CVPR 2006*, pages 11–18, 2006.

S. Nelson. Cortical microcircuits: diverse or canonical? *Neuron*, 36(1):19–27, 2002.

S. J. Nowlan and T. J. Sejnowski. A selection model for motion processing in area MT of primates. *Journal of Neuroscience*, 15(2):1195–214, 1995.

E. Oja. A simplified neuron model as a principal component analyzer. *J Math Biol*, 15(3): 267–73, 1982.

J. O'Keefe. Place units in the hippocampus of the freely moving rat. *Experimental Neurology*, 51(1):78–109, 1976.

A. Pasupathy and C. E. Connor. Responses to contour features in macaque area V4. *Journal of Neurophysiology*, 82:2490–2502, 1999.

A. Pasupathy and C. E. Connor. Shape representation in area V4: position-specific tuning for boundary conformation. *Journal of Neurophysiology*, 86(5):2505–19, 2001.

A. Pasupathy and C. E. Connor. Population coding of shape in area V4. *Nature Neuroscience*, 5(12):1332–1338, 2002.

D. I. Perrett and M. W. Oram. Neurophysiology of shape processing. *Image and Vision Computing*, 11(6):317–333, 1993.

T. Poggio. A theory of how the brain might work. *Cold Spring Harb Symp Quant Biol*, 55: 899–910, 1990.

T. Poggio and E. Bizzi. Generalization in vision and motor control. *Nature*, 431(7010): 768–74, 2004.

T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343(6255):263–6, 1990.

T. Poggio and F. Girosi. A theory of networks for approximation and learning. Technical Report AI Memo 1140, Massachusetts Institute of Technology, 1989.

D. A. Pollen, A. W. Przybyszewski, M. A. Rubin, and W. Foote. Spatial receptive field organization of macaque V4 neurons. *Cereb. Cortex*, 12(6):601–616, 2002.

N. J. Priebe, F. Mechler, M. Carandini, and D. Ferster. The contribution of spike threshold to the dichotomy of cortical simple and complex cells. *Nature Neuroscience*, 7(10):1113–22, 2004.

J. P. Rauschecker, B. Tian, and M. Hauser. Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*, 268(5207):111–4, 1995.

W. Reichardt, T. Poggio, and K. Hausen. Figure-ground discrimination by relative movement in the visual system of the fly. *Biological Cybernetics*, V46(0):1–30, 1983.

J. H. Reynolds, L. Chelazzi, and R. Desimone. Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience*, 19(5):1736–53, 1999.

M. Riesenhuber and T. Poggio. Are cortical models really bound by the "binding problem"? *Neuron*, 24(1):87–93, 111–25, 1999a.

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–25, 1999b.

D. L. Ringach. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology*, 88:455–463, 2002.

D. L. Ringach. Mapping receptive fields in primary visual cortex. *Journal of Physiology*, 558:717–728, 2004.

G. A. Rousselet, S. J. Thorpe, and M. Fabre-Thorpe. Taking the max from neuronal responses. *Trends Cogn Sci*, 7(3):99–102, 2003.

N. C. Rust, V. Mante, E. P. Simoncelli, and J. A. Movshon. How MT cells analyze the motion of visual patterns. *Nature Neuroscience*, 9(11):1421–1431, 2006.

S. Schein and R. Desimone. Spectral properties of V4 neurons in the macaque. *Journal of Neuroscience*, 10(10):3369–3389, 1990.

P. H. Schiller. Effect of lesions in visual cortical area V4 on the recognition of transformed objects. *Nature*, 376(6538):342–4, 1995.

P. H. Schiller and K. Lee. The role of the primate extrastriate area V4 in vision. *Science*, 251(4998):1251–3, 1991.

O. Schwartz and E. P. Simoncelli. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–25, 2001.

P. Series, J. Lorenceau, and Y. Fregnac. The "silent" surround of V1 receptive fields: theory and experiments. *Journal of Physiology - Paris*, 97(4-6):453–74, 2003.

T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. Technical Report CBCL Paper #259/AI Memo #2005-036, Massachusetts Institute of Technology, 2005.

T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. USA*, 104(15):6424–9, 2007a.

T. Serre and M. Riesenhuber. Realistic modeling of simple and complex cell tuning in the HMAX model, and implications for invariant object recognition in cortex. Technical Report CBCL Paper #239/AI Memo #2004-017, Massachusetts Institute of Technology, 2004.

T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans Pattern Anal Mach Intell*, 29(3):411–26, 2007b.

H. S. Seung. Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron*, 40(6):1063–73, 2003.

R. Shapley, M. Hawken, and D. L. Ringach. Dynamics of orientation selectivity in the primary visual cortex and the importance of cortical inhibition. *Neuron*, 38(5):689–99, 2003.

K. A. Sundberg, J. F. Mitchell, and J. H. Reynolds. Contrast dependant center-surround interactions in macaque area V4. *Journal of Vision*, 5(8):79–79, 2005.

K. Tanaka, H. Saito, Y. Fukuda, and M. Moriya. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, 66(1):170–189, 1991.

S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–2, 1996.

S. Thorpe and M. Imbert. Biological constraints on connectionist modelling. In *Connectionism in perspective*, pages 63–93. Elsevier, 1989.

V. Torre and T. Poggio. Synaptic mechanism possibly underlying directional selectivity to motion. *Proceedings of the Royal Society of London Series B-Biological Sciences*, 202 (1148):409–416, 1978.

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

W. E. Vinje and J. L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–6, 2000.

G. Wallis. Using spatio-temporal correlations to learn invariant object recognition. *Neural Netw*, 9(9):1513–1519, 1996.

G. Wallis and H. Bulthoff. Learning to recognize objects. *Trends Cogn Sci*, 3(1):22–31, 1999.

G. Wallis and E. T. Rolls. Invariant face and object recognition in the visual system. *Prog Neurobiol*, 51(2):167–94, 1997.

Y. Wang, I. Fujita, and Y. Murayama. Neuronal mechanisms of selectivity for object features revealed by blocking inhibition in inferotemporal cortex. *Nature Neuroscience*, 3(8):807–13, 2000.

H. R. Wilson and F. Wilkinson. Detection of global structure in Glass patterns: implications for form vision. *Vision Research*, 38(19):2933–47, 1998.

R. I. Wilson, G. C. Turner, and G. Laurent. Transformation of olfactory representations in the drosophila antennal lobe. *Science*, 303(5656):366–70, 2004.

L. Wiskott and T. J. Sejnowski. Slow feature analysis: unsupervised learning of invariances. *Neural Computation*, 14(4):715–70, 2002.

A. J. Yu, M. A. Giese, and T. A. Poggio. Biophysiologically plausible implementations of the maximum operation. *Neural Computation*, 14(12):2857–81, 2002.

A. L. Yuille and N. M. Grzywacz. A winner-take-all mechanism based on presynaptic inhibition feedback. *Neural Computation*, 1:334–347, 1989.

K. C. Zhang, M. I. Sereno, and M. E. Sereno. Emergence of position-independent detectors of sense of rotation and dilation with Hebbian learning - an analysis. *Neural Computation*, 5(4):597–612, 1993.

D. Zoccolan, D. D. Cox, and J. J. DiCarlo. Multiple object response normalization in monkey inferotemporal cortex. *Journal of Neuroscience*, 25(36):8150–64, 2005.

D. Zoccolan, M. Kouh, T. Poggio, and J. J. DiCarlo. Trade-off between shape selectivity and tolerance to identity-preserving transformations in moneky inferotemporal cortex. *In Submission*, 2007.