Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2006-028
CBCL-260

April 25, 2006

Learning a Dictionary of Shape-Components in Visual Cortex: Comparison with Neurons, Humans and Machines

Thomas Serre

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Learning a Dictionary of Shape-Components in Visual Cortex: Comparison with Neurons, Humans and Machines

by

Thomas Serre

Ingénieur de l'Ecole Nationale Supérieure
des Télécommunications de Bretagne, 2000
and
MS, Université de Rennes, 2000

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

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2006

© Massachusetts Institute of Technology 2006. All rights reserved.

Author .....
Department of Brain and Cognitive Sciences
April 24, 2006

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

Accepted by .....
Matt Wilson
Professor of Neurobiology and
Chairman, Department Graduate Committee

Learning a Dictionary of Shape-Components in Visual Cortex: Comparison with Neurons, Humans and Machines

by

Thomas Serre

Submitted to the Department of Brain and Cognitive Sciences
on April 24, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In this thesis, I describe a quantitative model that accounts for the circuits and computations of the feedforward path of the ventral stream of visual cortex. This model is consistent with a general theory of visual processing that extends the hierarchical model of [Hubel and Wiesel, 1959] from primary to extrastriate visual areas. It attempts to explain the first few hundred milliseconds of visual processing and "immediate recognition". One of the key elements in the approach is the learning of a generic dictionary of shape-components from V2 to IT, which provides an invariant representation to task-specific categorization circuits in higher brain areas. This vocabulary of shape-tuned units is learned in an unsupervised manner from natural images, and constitutes a large and redundant set of image features with different complexities and invariances. This theory significantly extends an earlier approach by [Riesenhuber and Poggio, 1999a] and builds upon several existing neurobiological models and conceptual proposals.

First, I present evidence to show that the model can duplicate the tuning properties of neurons in various brain areas (*e.g.*, V1, V4 and IT). In particular, the model agrees with data from V4 about the response of neurons to combinations of simple two-bar stimuli [Reynolds et al., 1999] (within the receptive field of the S_2 units) and some of the C_2 units in the model show a tuning for boundary conformations which is consistent with recordings from V4 [Pasupathy and Connor, 2001]. Second, I show that not only can the model duplicate the tuning properties of neurons in various brain areas when probed with artificial stimuli, but it can also handle the recognition of objects in the real-world, to the extent of competing with the best computer vision systems. Third, I describe a comparison between the performance of the model and the performance of human observers in a rapid animal *vs.* non-animal recognition task for which recognition is fast and cortical back-projections are likely to be inactive. Results indicate that the model predicts human performance extremely well when the delay between the stimulus and the mask is about $50\ ms$. This suggests that cortical back-projections may not play a significant role when the time interval is in this range, and the model may therefore provide a satisfactory description of the feedforward path.

Taken together, the evidences suggest that we may have the skeleton of a successful theory of visual cortex. In addition, this may be the first time that a neurobiological model, faithful to the physiology and the anatomy of visual cortex, not only competes with some

of the best computer vision systems thus providing a realistic alternative to engineered artificial vision systems, but also achieves performance close to that of humans in a categorization task involving complex natural images.

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

Acknowledgments

I acknowledge my advisor, Prof. Tomaso Poggio (McGovern Institute, MIT), for his mentoring during this thesis and for helping me transition from computer science to neuroscience. In particular, I would like to thank him for sharing his scientific thinking, shaping my own critical judgment and teaching me to always sort out important from irrelevant research results.

I acknowledge the members of my thesis committee Jim DiCarlo (McGovern Institute, MIT), Earl Miller (Picower Institute, MIT) and Simon Thorpe (CNRS, France) for their patience, guidance and advice.

I acknowledge my collaborators during the five years of this thesis work: First and foremost Tomaso Poggio (MIT) for his substantial contributions on all aspects of this work; Maximilian Riesenhuber (Georgetown) and Jennifer Louie (MIT) for their contributions in the early stages of this thesis; Charles Cadieu (Berkeley) and Minjoon Kouh for their contributions on the comparison between the tuning of model units and V4 data in Chapter 3; Stan Bileschi (MIT) and Lior Wolf (MIT) for their contributions in the comparison between the model and the computer vision systems in Chapter 4; Aude Oliva (MIT) for her contribution on the comparison between the model and human observers in Chapter 5; Rodrigo Sigala (Max Planck Institute) and Martin Giese (Tubingen) for their contributions in extending the present model of the ventral stream to the recognition of biological motion in the dorsal stream in Chapter 6; Dirk Walther (CalTech) and Christof Koch (CalTech) for their contributions in bringing together the present model of "pre-attentive" vision with a model of top-down attention in Chapter 6; Christof Koch (CalTech) and Gabriel Kreiman (MIT) for their contributions in proposing possible roles for the back-projections and in the model of mental imagery in Chapter 6.

I would like to thank my neuroscience fellow at CBCL, Charles Cadieu (Berkeley), Ulf Knoblich (MIT), Minjoon Kouh (MIT) and Gabriel Kreiman (MIT) for numerous discussions and various contributions in the development of the model. I would also like to thank several people for valuable discussions related to this work: Heinrich Bülthoff, Chou Hung, David Lowe, Pietro Perona, Antonio Torralba and Davide Zoccolan.

Many thanks to Yuri Ivanov now a friend after a summer internship under his supervision in the summer 2003 at Honda Research Inc. and Sanmay Das, my office mate and friend, for making my last 3 years at CBCL enjoyable.

To my wife Alison

Contents

1 Introduction **19**

A From Models to Theories 19

B Architecture and Function of the Visual Cortex 22

B.1 Hierarchical Organization 22

B.2 Learning and Plasticity 26

B.3 Feedforward Processing and Immediate Recognition 28

B.4 Summary 31

C Models of Object Recognition in Cortex 31

C.1 Related Work 32

C.2 Cortical Circuits and Key Computations 37

D Original Contributions 45

D.1 Learning a Dictionary of Shape-Components in Visual Cortex 45

D.2 Comparison with Neural Data 46

D.3 Comparison with Computer Vision Systems 46

D.4 Comparison with Human Observers 47

2 Theory and Basic Model Implementation **49**

A Building a Dictionary of Shape-Components from V1 to IT 50

A.1 Simple and Complex Cells in V1 50

A.2 Beyond V1: Features of Moderate Complexity 55

A.3 Invariant Recognition in IT 61

B Learning a Dictionary of Shape-Components from Natural Images 63

B.1 On Learning Correlations 63

B.2 The Learning Rule 63

C Building Task-Specific Circuits from IT to PFC 66
D Discussion 70
D.1 On Learning Good Features 70
D.2 One Basic Dictionary of Features for Different Recognition Tasks . . 72
D.3 What is the Other 99% of Visual Cortex Doing? 74

3 Comparison with Neurons **81**
A V1 and the Model 81
A.1 Methods 82
A.2 Results 84
B V4 and the Model 85
B.1 Tuning to Boundary-Conformation Stimuli 85
B.2 Tuning to Two Bar-Stimuli 91
B.3 Results 91
C TE and the Model 94
C.1 Methods 94
C.2 Results 95
D Discussion 95

4 Performance on Natural Images **99**
A Robustness of the Model 100
A.1 Performance on the CalTech-101 Database 100
A.2 Approximating the Key Computations 105
B Comparison with Standard AI Recognition Systems 109
C Discussion 111

5 Predicting Human Performance in a Rapid Categorization Task **117**
A Introduction 118
B Experiment 1 123
C Experiment 2 124
D Experiment 3 127
E Methods 127
E.1 Model architecture 127

E.2 Classifier from IT to PFC 129
F Supplementary Information 129
F.1 Supplementary Methods 129
F.2 On Interrupting Back-Projections with the Mask 133
F.3 Supplementary Data 134

6 Discussion **139**
A Summary 139
B Open Questions 140
B.1 The Architecture 140
B.2 Learning and Plasticity 141
B.3 Performance on Natural Images 141
C Future Extensions 142
C.1 The Ventral Pathway 142
C.2 The Dorsal Pathway 142
C.3 Including Back-Projections 143
D Predictions 148
E Beyond Vision: A Universal Scheme? 149

A Detailed Model Implementation and Parameters **153**
A Model Architecture and Implementation 153
B Major Extensions from the Original HMAX model 160

B Additional Comparisons with Computer Vision Systems **163**
A Image Datasets 163
B Results 164
C Summary of All Comparisons with Computer Vision Systems 171

List of Figures

1-1 The ventral stream of visual cortex and object recognition 23
1-2 The organization of visual cortex 24
1-3 The feedforward circuits involved in rapid categorization tasks 29
1-4 The Hubel & Wiesel hierarchical model for building complex cells from simple cells 33
1-5 Computational modules in cortex 38
1-6 Bell-shaped TUNING from one AIT cell 39
1-7 MAX operation from a complex cell in area 17 of the cat 41
1-8 A possible cortical circuit for TUNING and MAX operations 45

2-1 Architecture of the model 51
2-2 Receptive field organization of the S_1 units 52
2-3 Functional organization in the model 53
2-4 Orientation tuning at the S_1 level 54
2-5 Building tolerance to scale and position at the C_1 level 55
2-6 Building S_2 and C_2 units 56
2-7 Sample natural images used to expose the model and learn the generic dictionary of shape-components from V2 to IT 65
2-8 Good *vs.* bad feature representations 71
2-9 How learning a good representation improves classification performance . 71
2-10 Expert *vs.* generic features 73

3-1 Spatial frequency bandwidth of the C_1 *vs.* S_1 units 84
3-2 The stimulus dataset used to measure the tuning of cells to boundary conformations 86

3-3 Tuning for boundary conformation: Model C_2 units *vs.* V4 neurons (I) . . . 89
3-4 Tuning for boundary conformation: Model C_2 units *vs.* V4 neurons (II) . . . 90
3-5 Tuning for two-bar stimuli: Experimental procedure 92
3-6 Tuning for two-bar stimuli: Model C_2 units *vs.* V4 neurons 93
3-7 Paperclip stimuli used to test the tuning and invariance properties of the model S_4 units 96

4-1 Representative images from the *CalTech-101* object database 100
4-2 Schematic of the "simplified" architecture used to compare the model to other benchmark AI recognition systems 102
4-3 Sample results from the *CalTech-101* object database 104
4-4 Robustness of the model to approximations in the key computations (I) . . . 107
4-5 Robustness of the model to approximations in the key computations (II) . . 107
4-6 The *CalTech-5* datasets 108
4-7 The *MIT-CBCL* datasets 109

5-1 Schematic of the model implementation used in the comparison with human-observers on the animal *vs.* non-animal categorization task 121
5-2 Schematic of the animal *vs.* non-animal categorization task 122
5-3 Results of experiment 1 125
5-4 Results of experiment 2 126
5-5 Results of experiment 3 128
5-6 Estimate of the timing of feedback loops in the ventral stream of primate visual cortex 134

6-1 An illustration of the "focused" classifier. 145
6-2 A simple model of mental imagery. 147
6-3 The same S_1 type of receptive field structures can produce different measurements when placed along different visual directions. 150
6-4 S_1 units in the auditory system. 150

B-1 Model (C_{2b} unit responses) *vs.* SIFT features 165
B-2 Performance *vs.* size of the dictionary of C_{2b} units and on the number of training examples 166

B-3 Overall performance on the *CalTech-101* 168
B-4 Sample results by the StreetScene recognition system 170

List of Tables

2.1 Number of units in the model 75

3.1 Summary of the tuning properties of the S_1 and C_1 units *vs.* parafoveal simple and complex cells from monkey primary visual cortex 83

3.2 Goodness of fit of different tuning functions for model units and V4 neurons 88

4.1 A comparison between the model and other AI benchmark recognition systems 111

5.1 Summary of mean reaction times and standard error means for human observers 135

5.2 Summary of reaction times (10^{th} percentile) for human observers 136

5.3 Summary of accuracy measures for human observers and the model (I) . . . 137

5.4 Summary of accuracy measures for human observers and the model (II) . . 138

6.1 Mental imagery in the model 147

A.1 Summary of all the model parameters 159

B.1 Summary of all the comparisons performed between the model and other computer vision systems 173

Chapter 1

Introduction

A From Models to Theories

Since the 50's and the groundbreaking work of Hodgkin & Huxley to model the spike generation process [Hodgkin and Huxley, 1952], there has been an explosion in the development of computational models for neuroscience. By now, there are probably hundreds of models for early vision alone. Indeed some of them have been quiet successful in characterizing the early visual processing from the retina through LGN and V1 (see [Carandini et al., 2005]).

Perhaps one of the most influential model in vision is the Reichardt model of motion detection [Reichardt, 1961] which began as a model of the optomotor response of the beetle and later influenced work on motion in different species [Barlow and Lewick, 1965; Egelhaaf and Reichardt, 1987; Borst et al., 2005] and even human psychophysics [Adelson and Bergen, 1985]. Similarly the gain control model [Heeger, 1992a,b] has been shown to account for a wide array of visual phenomena both at the level of single cortical neuron responses (*e.g.*, luminance and gain control, contrast adaptation, surround suppression and contextual effects as well as orientation tuning and motion selectivity [Heeger, 1993; Carandini and Heeger, 1994; Heeger et al., 1996; Tolhurst and Heeger, 1997]) and psychophysics [Watson and Solomon, 1997]. Also in primary visual cortex, very detailed simulations [Ben-Yishai et al., 1995; Somers et al., 1995; McLaughlin et al., 2000] (see [Lund et al., 2003] for a review) of small networks of neurons (≈ 1 mm^2 of cortex) are contributing towards the understanding of the mechanisms for orientation selectivity.

Detailed models of higher cortical areas, however, have been more scarce. For instance,

a model of the cortical circuits between V1 and V2 has been described in [Raizada and Grossberg, 2001] and a two-stage model of MT responses by [Simoncelli and Heeger, 1998]. Surprisingly there have been relatively few attempts to address a high-level computational task, *e.g.*, flexible control by prefrontal cortex [Rougier and Reilly, 2002; Rougier et al., 2005] or probabilistic (Bayesian) models of reasoning and inference [Knill and Richards, 1996]. The latter have been particularly useful for interpreting psychophysical experiments and constrain theories of perception, see [Knill and Richards, 1996; Mamassian et al., 2002; Rao et al., 2002; Kersten and Yuille, 2003]. For instance, Weiss *et al.* showed that the same ideal observer model can explain numerous illusions thought to be mediated by different neural mechanisms [Weiss et al., 2002]. Yet such probabilistic models lack explicit correspondences between functional primitives of the model and structural primitives of the cortex and their implications in helping understand neural processing are at best only indirect.

Altogether beyond biologically-inspired algorithms [Fukushima, 1980; LeCun et al., 1998; Ullman et al., 2002; Wersing and Koerner, 2003], *i.e.*, systems only qualitatively constrained by the anatomy and physiology of the visual cortex, there have been very few neurobiologically plausible models [Perrett and Oram, 1993; Mel, 1997; Wallis and Rolls, 1997; Riesenhuber and Poggio, 1999a; Thorpe, 2002; Amit and Mascaro, 2003], that try to address a generic, high-level computational function such as object recognition by summarizing and integrating a large body of data from different levels of understanding. To paraphrase C.F Stevens, *"Models are common; good theories are scarce"* [Stevens, 2000].

The past decades of work in striate and extrastriate cortical areas have produced a significant and rapidly increasing amount of data. Understanding the mechanisms underlying object recognition will require to bridge the gap between several levels of understanding from the information processing or computational level to the level of circuits and of cellular and biophysical mechanisms. The emerging picture of how cortex performs object recognition is in fact becoming too complex for any simple, qualitative "mental" model. There is a need for quantitative computational theories that could 1) summarize and organize existing data and 2) help planning, coordinating and interpreting new experiments. With the advent of supercomputers and dedicated architectures to simulate detailed models of neural processing, *e.g.*, *blue brain*[1] [Markram, 2006], it will soon be possible to simulate detailed circuits of the cortex organized across several cortical areas (*e.g.*, part of the

visual system).

In particular, one of the main challenges in understanding the mechanisms of object recognition in cortex is to determine the selectivity of individual neurons. Linear techniques such as *reverse-correlation* (see [Dayan and Abbott, 2001]) have shown some success in characterizing V1 simple cell receptive fields. Because of the non-linear nature of the neural responses in extrastriate areas, it is clear that, without a prior model, any systematic method is doomed to fail [Albright and Gross, 1990]. Indeed methods for characterizing the neural response beyond V1 have remained either ad hoc, with an experimenter presenting random objects (*e.g.*, faces, hands, or toilet brushes) or subjective. For instance, the feature reduction method by Tanaka and colleagues [Tanaka, 1996] have been shown to be quiet misleading in some cases [Knoblich and Riesenhuber, 2002].

By increasing the contribution of computational models, Neuroscience is slowly making up its lag behind other experimental sciences such as Physics. Yet the acceptance of models in Neuroscience is not unanimous. Partly, this may come from the fact that, so far, models have only been applied to "simplistic" problems such as the recognition of artificial "idealized" objects – which are typically used in psychology and physiology (*e.g.*, paperclips on a blank background [Logothetis et al., 1994, 1995; Riesenhuber and Poggio, 1999a]). In particular, the ability of these biologically plausible models to explain more natural real-world scenarios (*i.e.*, unsegmented objects in clutter undergoing large variations in shape, pose, appearance and illumination conditions), have been questioned. In this thesis, we take on the challenge and describe a quantitative theory of object recognition in primate visual cortex that 1) bridges several levels of analysis from biophysics and physiology to behavior and 2) achieves human level performance on the rapid recognition of complex natural images. The theory is restricted to the feedforward path of the ventral stream and therefore to the first 150 milliseconds or so of visual recognition. The theory evolved over the past years as a result of a collaboration between several theorists and experimentalists. In particular, the theory significantly extends an earlier model [Riesenhuber and Poggio, 1999a, 2000]. In this thesis, I shall emphasize my own contributions (see Section D) but for a further overview of the theory please refer to [Serre et al., 2005a].

We start by describing some general knowledge and basic facts about the ventral stream of primate visual cortex in Section B. We then review work on models of object recognition in cortex in Section C. Finally we list the contributions from this thesis in Section D.

B Architecture and Function of the Visual Cortex

The visual cortex is composed of several areas that tend to be hierarchically organized [Felleman and van Essen, 1991] (see Fig. 1-1). It is generally believed that the flow of information through visual cortex can be dissociated into two streams [Mishkin et al., 1983; DeYoe and Essen, 1988] : the *ventral* stream and the *dorsal* stream. Object recognition in cortex is thought to be mediated by the ventral visual pathway [Ungerleider and Haxby, 1994; Tanaka, 1996; Logothetis and Sheinberg, 1996] which is organized into a series of neurally interconnected stages, starting from the retina, through the Lateral Geniculate Nucleus (LGN) of the thalamus to primary visual cortex (V1) and extrastriate visual areas, V2, V4 and IT. It, in turn, is believed to play a key role in invariant object recognition [Tanaka, 1996] and provides a major source of input to prefrontal cortex (PFC) involved in linking perception to memory and action [Miller, 2000].

Over the last decades, several physiological studies in non-human primates have established a core of basic facts about cortical mechanisms of recognition that seem to be widely accepted and that confirm and refine older data from neuropsychology. Fig. 1-2 (modified from [Perrett and Oram, 1993]) illustrate these general, mostly accepted, properties of the feedforward path of the ventral stream architecture.

B.1 Hierarchical Organization

Building an Invariant Representation from V1 to IT

There is now a large body of evidences that suggest a gradual increase in both the invariance properties and the complexity of the preferred stimuli of neurons along the visual stream. The notion of a hierarchy of visual processing initiated with the groundbreaking work of Hubel & Wiesel, first in the primary visual cortex of the cat [Hubel and Wiesel, 1959, 1962, 1965] and then of the macaque [Hubel and Wiesel, 1968, 1977]. In particular they described how, from the arrangement of several simple cells with small receptive fields that respond best to a bar at a particular orientation and position, a complex cell response can be obtained, that respond also to a bar at a particular orientation anywhere within its receptive field. Beyond V1, neurons along the ventral stream show an increase in the size of their receptive fields as well as in the complexity of their preferred stimuli [Perrett and Oram, 1993; Kobatake and Tanaka, 1994; Tanaka, 1996; Logothetis and Shein-

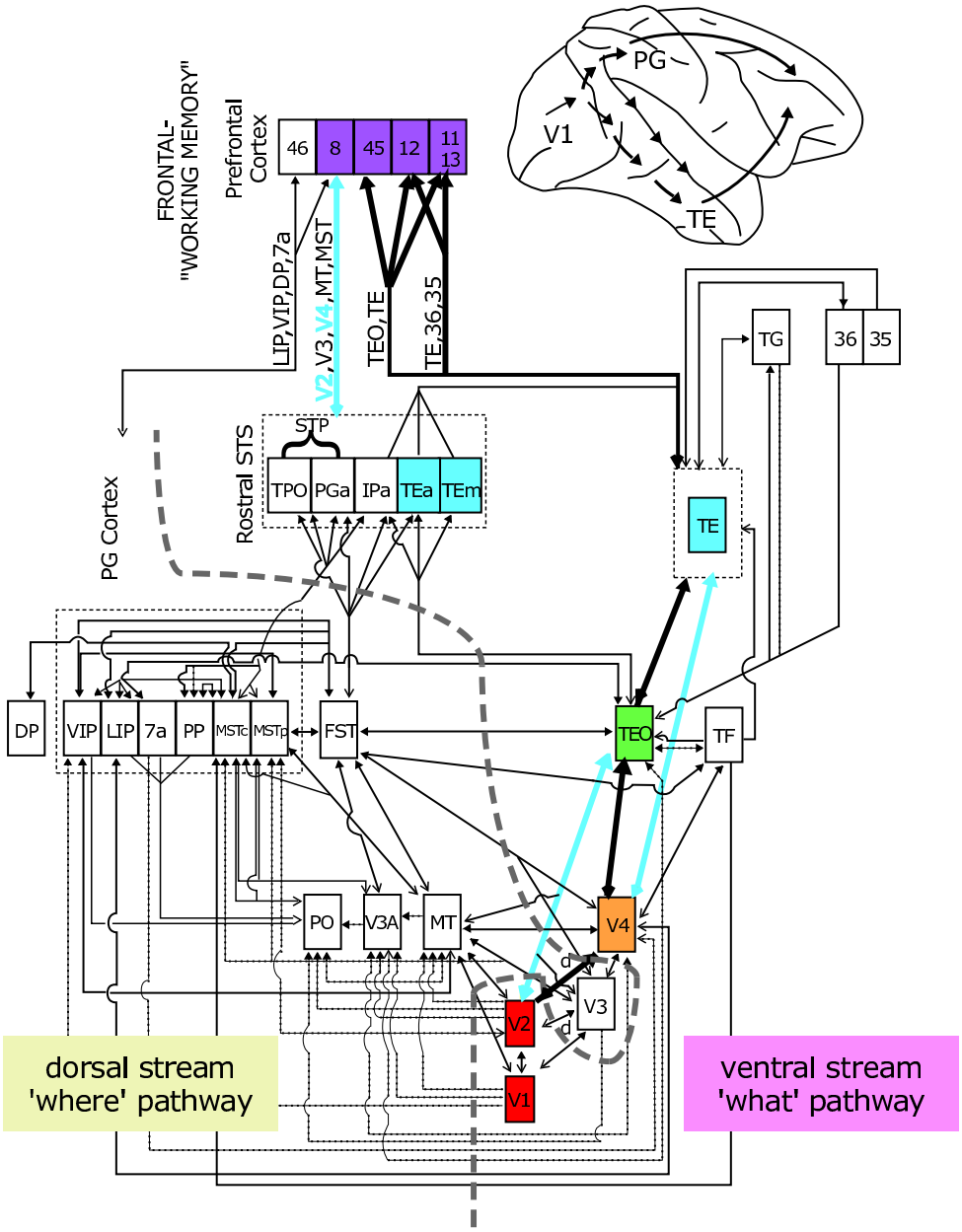


Figure 1-1: The ventral stream of visual cortex and object recognition. Modified from Ungerleider & Van Essen [Gross, 1998].

berg, 1996; Riesenhuber and Poggio, 2002]. For instance in V2, it has been shown that some neurons respond to angle stimuli, possibly through the non-linear combination of oriented subunits [Boynton and Hegdé, 2004]. Further along the hierarchy, neurons in V4 have been shown to respond to object features of moderate complexity [Kobatake et al., 1998], such as Cartesian and non-Cartesian grating [Gallant et al., 1996] or the combination of boundary-conformations [Pasupathy and Connor, 1999, 2001, 2002].

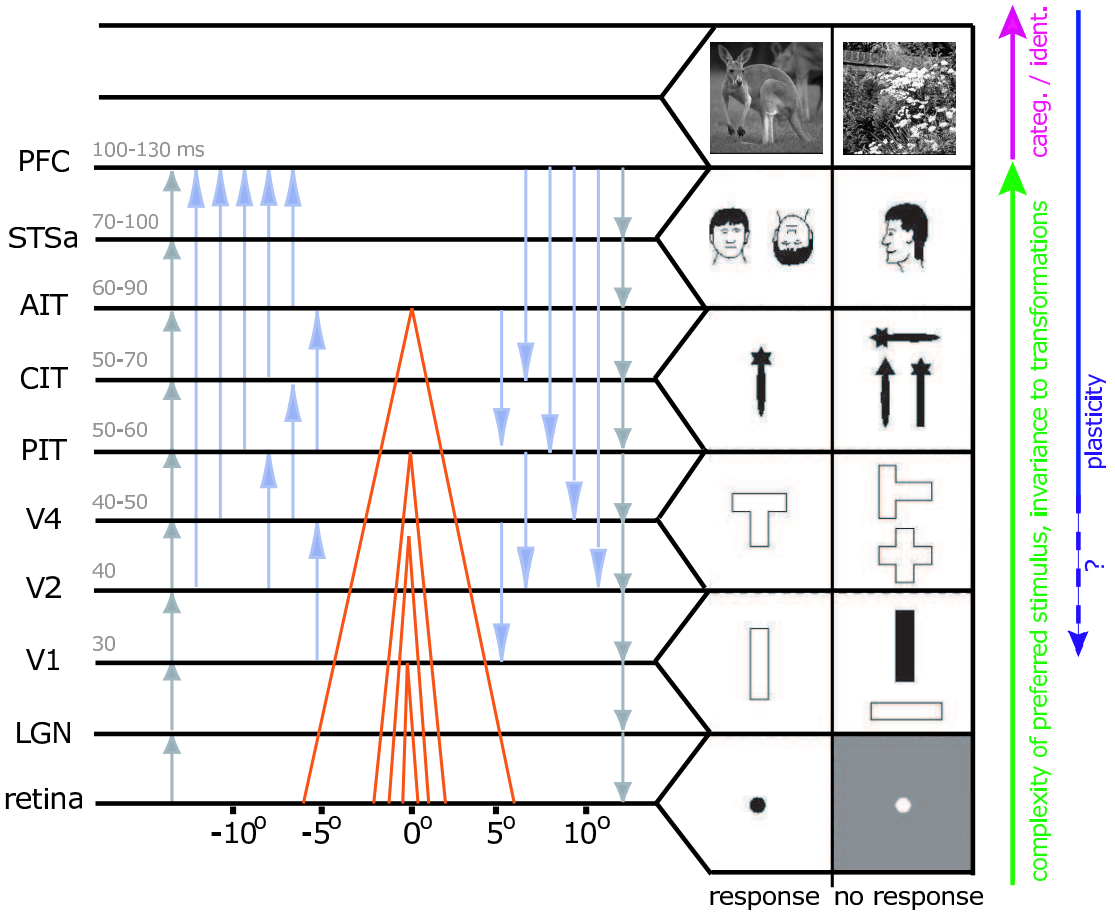


Figure 1-2: The organization of visual cortex based on a core of knowledge that has been accumulated over the past 30 years. The figure is modified from [Oram and Perrett, 1994] mostly to include the likely involvement of prefrontal cortex during recognition tasks by setting task-specific circuits to read-out shape information from IT [Scalaidhe et al., 1999; Freedman et al., 2002, 2003; Hung et al., 2005].

Beyond V4, in IT, many neurons are selective for a variety of stimulus attributes, such as color, orientation, texture, direction of movement, and the vast majority is tuned to various shapes [Gross et al., 1972; Desimone and Gross, 1979; Desimone et al., 1984; Logothetis et al., 1995; Tanaka, 1996; Logothetis and Sheinberg, 1996; Vogels, 1999; op de Beeck et al., 2001; Brincat and Connor, 2004]. At the top of the ventral stream, in anterior inferotemporal cortex (AIT), cells are found that are tuned to complex stimuli including body parts, *e.g.*, faces and face parts, hands, as well as other body parts [Gross et al., 1972; Bruce et al., 1981; Perrett et al., 1982; Rolls, 1984; Perrett et al., 1984; Baylis et al., 1985; Perrett et al., 1987; Yamane et al., 1988; Hasselmo et al., 1989; Perrett et al., 1991, 1992; Hietanen et al., 1992; Souza et al., 2005] (see [Logothetis and Sheinberg, 1996] for a review).

A hallmark of these AIT cells is the robustness of their firing to stimulus transformations such as scale and position changes [Tanaka, 1996; Logothetis and Sheinberg, 1996; Logothetis et al., 1995; Perrett and Oram, 1993]. In addition, as other studies have shown [Perrett et al., 1985; Perrett and Oram, 1993; Booth and Rolls, 1998; Logothetis et al., 1995; Hietanen et al., 1992], most neurons show specificity for a certain object view or lighting condition. In particular, Logothetis *et al.* trained monkeys to perform an object recognition task with isolated views of novel 3D objects (*e.g.*, paperclips) [Logothetis et al., 1995]. When recording from the animals' IT, they found that the great majority of neurons selectively tuned to the training objects were view-tuned (with a half-width of about 20^o for rotation in depth) to one of the training objects (about one tenth of the tuned neurons were view-invariant, in agreement with earlier predictions [Poggio and Edelman, 1990]). Interestingly they also found that, while monkeys were trained with the object at the same retinal location and size, neurons *naturally* exhibited an average translation invariance of $\sim 4^o$ (for typical stimulus sizes of 2^o) and an average scale invariance of two octaves [Riesenhuber and Poggio, 1999a]. Whereas view-invariant recognition requires visual experience of the specific novel object, significant position and scale invariance seems to be immediately present in the view-tuned neurons [Logothetis et al., 1995] without the need of visual experience for views *of the specific object* at different positions and scales.

Beyond IT: Task-Specific Circuits

The results by Logothetis *et al.* are in agreement with a general computational theory [Poggio, 1990; Poggio and Girosi, 1990; Poggio and Edelman, 1990; Poggio and Hurlbert, 1994; Vetter et al., 1995; Riesenhuber and Poggio, 2000] suggesting that a variety of visual object recognition tasks (involving the categorization of objects and faces at different levels) can be performed based on a linear combination of a few units tuned to specific task-related training examples. From a computational perspective, contrary to affine transformations such as translation and rescaling, invariances to non-affine transformations such as illumination, pose, *etc* require specific examples from the target object undergoing the desired transformation. This suggested, in agreement with the physiology (see above) that a majority of neurons in IT should exhibit a range of invariance to changes in position and scale, yet, be highly sensitive to changes in 3D rotation and illumination.

As suggested by [Riesenhuber and Poggio, 2000], a generic dictionary of shape-components, from V1 to IT, may provide position and scale-invariant inputs to task specific circuits beyond IT to generalize over non-affine transformations. For instance, pose-invariant face categorization circuits may be built, possibly in PFC, by combining several units tuned to different face examples, including different people, views and lighting conditions. "Animal *vs.* non-animal" categorization units could be built by combining the activity of a few AIT cells tuned to various examples of animals and non-animals. A study by [Freedman et al., 2003] recently suggested that the tuning of neurons in IT is best explained by their selectivity to shape while the tuning of neurons in PFC is best explained by their selectivity to object category. While it is often difficult to tell apart shape-selectivity from category-selectivity (see [Freedman et al., 2003]), category-selectivity does not need to correspond to shape similarity. While tuning for shape can be learned in an unsupervised manner, category-specific tuning requires supervision (*i.e.*, training examples along with a corresponding label).

B.2 Learning and Plasticity

There is now good evidence for learning and plasticity in adult cortex. From the computational perspective, it is very likely that learning may occur in all stages of visual cortex. For instance if learning a new task involves high-level object-based representations, learning is likely to occur high-up in the hierarchy, at the level of IT or PFC. Conversely, if the task to be learned involves the fine discrimination of orientations, changes are more likely to occur in lower areas at the level of V1, V2 or V4. It is also very likely that changes in higher cortical areas should occur at faster time scales than changes in lower areas.

There have been several reports of plasticity at the level of PFC [Rainer and Miller, 2000; Freedman et al., 2003; Pasupathy and Miller, 2005]. It has also been shown [Miyashita, 1988; Sakai and Miyashita, 1991] that after training animals to perform delayed-match-to-sample tasks, some neurons in IT become selective to both the sample and the test stimuli while others become selective for the target stimulus during the delay period. The former is compatible with plasticity occurring at the level of IT while the latter suggests that changes occurred in higher stages, possibly in the medial temporal lobe or PFC. Such learning-related effects can be very fast: [Erickson and Desimone, 1999] reported that it may take as little as two days for them to occur.

Numerous studies have confirmed that the tuning of the view-tuned and object-tuned cells in AIT depends on visual experience and that neurons tend to be more selective for familiar than unfamiliar objects [Li et al., 1993; Booth and Rolls, 1998; Vogels, 1999; DiCarlo and Maunsell, 2000; Sheinberg and Logothetis, 2001; Freedman et al., 2003, 2006] or geometric shapes [Miyashita, 1993; Sakai and Miyashita, 1994; Logothetis et al., 1995; Tanaka, 1996; Kobatake et al., 1998; Miyashita and Hayashi, 2000; Baker et al., 2002; Sigala and Logothetis, 2002; op de Beeck et al., 2003]. In particular, it has been shown that extrinsic factors such as repetition, familiarity, and saliency can modulate the activity of IT neurons [Miller et al., 1991, 1993; Fahy et al., 1993; Li et al., 1993; Jagadeesh et al., 2001], and that visual experience results in increased clustering of neurons that respond selectively to trained stimuli [Erickson et al., 2000].

In addition, long-term visual experience and training have been shown to induce learning-related changes in IT. In particular, [Logothetis et al., 1995] showed that after training monkeys to discriminate between new unfamiliar objects (*e.g.*, paperclips), some AIT neurons become selective to particular views. [Kobatake et al., 1998] more directly showed that the population of cells selective for training examples was significantly higher (25%) in trained than in (untrained) control animals. [Sigala and Logothetis, 2002] found an enhanced representation of shape features that are relevant for categorizing sets of familiar stimuli and [Baker et al., 2002] for conjunctions of familiar stimulus feature pairs that are experienced together. [Booth and Rolls, 1998] showed that training is not necessary and that passive exposure to new 3D objects (real toy objects disposed in the monkey cage) is sufficient to produce view-dependent as well as view-independent neurons in IT that are selective for the target object. These results were recently confirmed by [Freedman et al., 2006] who additionally reported that this sharpening of the selectivity for the familiar objects is particularly pronounced during the early response onset of the neurons. Finally, imaging studies [Dolan et al., 1997; Gauthier et al., 1999] have shown an enhanced activity in IT during perceptual learning of objects and faces [Gauthier et al., 1999].

In intermediate cortical stages, at the level of V4, two studies have reported changes associated with perceptual learning on degraded images at the level of V4 [Yang and Maunsell, 2004; Rainer et al., 2004]. Below V4, learning-related changes have been reported in V1 [Singer et al., 1982; Karni and Sagi, 1991; Yao and Dan, 2001; Schuett et al., 2001; Crist et al., 2001], although their extent and functional significance is still under debate [Schoups et al., 2001; Ghose et al., 2002; DeAngelis et al., 1995].

B.3 Feedforward Processing and Immediate Recognition

Behavioral studies: It is well known that recognition is possible for scenes viewed in rapid visual serial presentations (RSVP) that do not allow sufficient time for eye movements or shifts of attention [Potter, 1975, 1976] (see also [Biederman, 1972; Biederman et al., 1974] and [Potter et al., 2002] for a recent review). In particular, [Potter, 1975, 1976] showed that human observers can detect a target object embedded in an image sequence when presented at rates as fast as $10/s$. In a pioneering series of experiments, Thorpe and colleagues introduced the study of a visual phenomenon referred to as *ultra-rapid visual categorization* [Thorpe et al., 1996] or simply *rapid categorization*. Over the years, several key characteristics of these rapid categorization tasks have been discovered. Below is a short overview:

1. Not only human observers, but also monkeys can be very fast and accurate during rapid categorization tasks. While slightly less accurate, monkeys are indeed $\sim 30\%$ faster than humans [Fabre-Thorpe et al., 1998].

2. Rapid categorization is not only possible for *natural* categories such as animals or food [Thorpe et al., 1996; Fabre-Thorpe et al., 1998] but also *artificial* categories such as means of transport [VanRullen and Thorpe, 2001b].

3. The removal of color information during image presentations has little effect on performance, leaving the latencies of the fastest behavioral responses unaffected in both monkeys and humans [Delorme et al., 2000].

4. The fastest reaction times cannot be further speed up by training and familiarity [Fabre-Thorpe et al., 2003].

5. Rapid categorization is possible even without direct fixation, *i.e.*, when presentations appear both near and far from the fovea [Thorpe et al., 2001b].

6. Rapid categorization is very robust to image rotation [Rousselet et al., 2003; Guyonneau et al., 2005] in terms of both reaction times and performance.

7. Rapid categorization is possible with presentation times as low as $6.25\ ms$ and when a backward mask follows the image presentation. Performance is near optimal for a stimulus onset asynchrony (*i.e.*, the delay between the stimulus and the mask) around $40 - 50\ ms$ [Bacon-Mace et al., 2005].

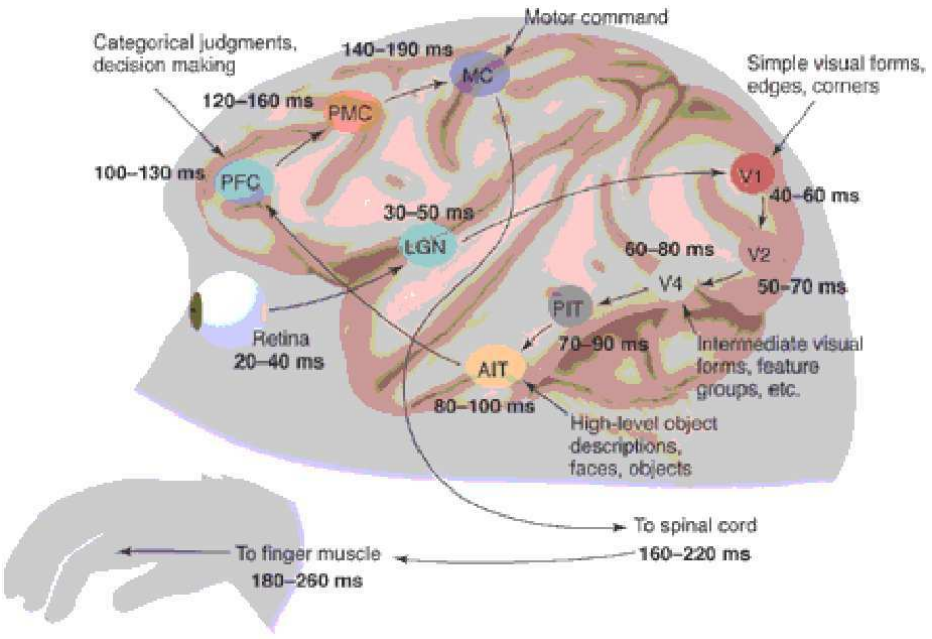


Figure 1-3: The feedforward circuits involved in rapid categorization tasks. Numbers for each cortical stage corresponds to the shortest latencies observed and the more typical mean latencies [Nowak and Bullier, 1997; Thorpe and Fabre-Thorpe, 2001]. Modified from [Thorpe and Fabre-Thorpe, 2001].

8. Rapid categorization does not seem to require attention. The level of performance of human observers remain high even when two images are flashed simultaneously (one on each hemifield) [Rousselet et al., 2002, 2004b] and when the image is presented parafoveally while an attention-demanding (letter discrimination) task is performed at the fovea [Li et al., 2002].

9. Differential EEG activity suggests that the task is solved within $150\ ms$ [Thorpe et al., 1996; Liu et al., 2002; Mouchetant-Rostaing et al., 2000] (but see also [Johnson and Olshausen, 2003; VanRullen and Thorpe, 2001a]).

10. Rapid categorization has also been studied using a choice saccade task. Indeed participants can make a saccade towards one of two images (flashed simultaneously for $30\ ms$ in each hemifield) that contains an animal with the most rapid saccades occurring within $150\ ms$ [Kirchner and Thorpe, 2005].

Altogether, considering typical neural latencies and the number of cortical stages involved in object categorization (see Fig. 1-3 for illustration), the very short reaction times observed during rapid categorization tasks strongly suggest that the flow of information

is mostly feedforward (apart from local feedback loops) and that there is no time for more than a few spikes at each stage of the hierarchy [Thorpe and Fabre-Thorpe, 2001] (see also [VanRullen and Koch, 2003]).

Physiological studies: At the neural level, the immediate selectivity of neurons after response onset is likely to rule out the involvement of feedback loops. [Oram and Perrett, 1992] showed that the response in IT neurons begins $80-100\ ms$ after onset of the visual stimulus and the response is tuned to the stimulus essentially from the very beginning. Indeed [Tovee et al., 1993] showed that $20-50\ ms$ time periods are sufficient to provide reasonable estimates of the firing rate and that the first $50\ ms$ after the onset of the neural response already contains 84% of the information present in a $400\ ms$ window. [Keysers et al., 2001] used a rapid serial visual presentation (RSVP) paradigm to assess the selectivity of neurons in STS and confirmed that stimulus discrimination can arise within $10-20\ ms$ of response onset, see also [Rolls et al., 1999; Ringach et al., 1997; Celebrini et al., 1993; Oram and Perrett, 1992; Thorpe et al., 1996]. Recent data [Hung et al., 2005] show that the activity of small neuronal populations (≈ 100 randomly selected cells) in IT over very short time intervals (as small as $12.5\ ms$ but lasting at least $50\ ms$) after beginning of the neural response ($80-100\ ms$ after onset of the stimulus) contains surprisingly accurate and robust information supporting a variety of recognition tasks.

Altogether, it has been suggested that for *immediate recognition* tasks, only a few spikes are propagated from one layer to the next [Thorpe and Imbert, 1989; Oram and Perrett, 1992; Tovee et al., 1993] and that the underlying architecture has to be feedforward (besides local recurrent loops to implement key computations). As suggested by [Földiák and Young, 1995; Perrett et al., 1998; Keysers et al., 2001], with only very few spikes transmitted at each stage, reading out information from one stage by the next is not about how much one neuron fires but rather how many of a particular type fire.

Anatomical studies: Studies of cortico-cortical circuits (*e.g.*, from V1 to extrastriate areas) have shown that feedforward connections are focused, while feedback connections (*e.g.*, from extrastriate cortex to V1) are more widespread [Callaway, 1998a; Zeki and Shipp, 1985; Shipp and Zeki, 1989a,b; Salin and Bullier, 1995]. Yet, despite the widespread nature of feedback connections, classical receptive fields (in V1 for instance) are relatively

small. This suggests that feedforward inputs shape the selectivity of individual neurons while feedback connections play a modulatory role, influencing neuronal responses primarily when visual stimuli are placed outside the classical receptive field (see [Knierim and van Essen, 1992; Bullier et al., 1996] for instance). Back-projections are neither sufficient (*i.e.,* they can't activate their target neurons without feedforward inputs [Zeki and Shipp, 1988; Sillito et al., 1994], see [Grossberg, 2005] for a review), nor necessary (neurons tend to be selective from the very beginning of the onset of their responses before back-projections could be active). Indeed, one criteria often used to isolate back-projections is that they are only activated after the neuron onto they project [Callaway, 1998a].

B.4 Summary

The accumulated evidence points to several mostly accepted properties of the ventral stream of visual cortex:

1. Along the hierarchy, neurons become both increasingly selective to more and more complex stimuli and increasingly invariant; first to 2D affine transformations (*e.g.,* position and scale, from V1 to IT) and then more complex transformations that require learning (*e.g.,* pose, illumination, *etc* , above IT). In parallel, the size of the receptive fields of neurons increase;

2. Learning can induce fast changes (within days) on the tuning properties of neurons probably at all stages and certainly in higher areas (IT and PFC).

3. The processes that mediate *immediate recognition* are likely to be feedforward, do not involve color information nor attentional circuits.

C Models of Object Recognition in Cortex

Models that have been proposed to explain invariant recognition in cortex roughly fall into two categories: the *normalization* approach and the *full replication* scheme (also referred to as invariant feature or convolutional networks), see [Riesenhuber and Poggio, 1999a; Ullman and Soloviev, 1999; Wiskott, 2006] for reviews.

The normalization approach is the standard approach in computer vision: Typically an input image is first transformed into an image-pyramid before it is scanned over all

positions and scales with a fixed size template window (see [Sung and Poggio, 1998; Osuna et al., 1997; Oren et al., 1997; Schneiderman and Kanade, 2000; Heisele et al., 2001b,c; Viola and Jones, 2001] to name just a few). A biologically plausible implementation of such normalization scheme is the *shifter-circuit* [Olshausen et al., 1993] and its extension [Olshausen et al., 1995] (see also [Postma et al., 1997]). In their approach, dynamic routing circuits control the connection strengths between input and output layers (switching on and off connections) so as to extract a normalized representation in the attended region. Related and perhaps more plausible models such as the *Gain-field* models have also been proposed [Salinas and Abbott, 1997; Riesenhuber and Dayan, 1997] that rely on attention-controlled shift or modulation of receptive fields in space.

All these models rely heavily on back-projections and top-down mechanisms. While it is possible that similar mechanisms may be used in visual cortex (for instance the gain-field models receive partial support from V4 data [Moran and Desimone, 1985; Connor et al., 1997]), it is clear that such circuits are not compatible with the physiological constrains provided by immediate recognition and rapid categorization tasks. Such circuits may be very important for normal everyday vision; yet, as discussed in Section B.3, there is now a large body of evidence suggesting that back-projections do not play a key role in the first few hundreds of milliseconds of visual processing. We now briefly review the literature on feedforward models of object recognition in cortex.

C.1 Related Work

Conceptual proposals: Following their work on striate cortex, Hubel & Wiesel proposed a hierarchical model of cortical organization. In particular, they described a hierarchy of cells within the primary visual cortex: at the bottom of the hierarchy, the *radially symmetric* cells are like LGN cells and respond best to small spots of light. Second, the *simple* cells do not respond well to spots of light and require bar-like (or edge-like) stimuli at a particular orientation, position and phase (*i.e.*, white bar on a black background or dark bar on a white background). In turn, the *complex* cells are also selective for bars at a particular orientation but they are insensitive to both the location and the phase of the bar within their receptive fields. At the top of the hierarchy the *hypercomplex* cells not only respond to bars in a position and phase invariant way, just like complex cells, but are also selective for bars of a particular length (beyond a certain length their response starts decreasing).

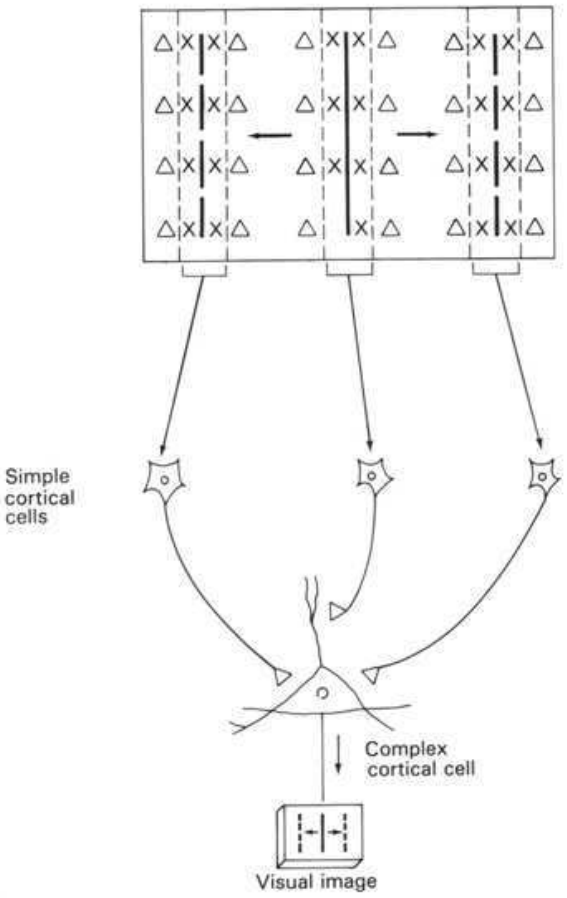


Figure 1-4: The Hubel & Wiesel hierarchical model for building complex cells from simple cells. Reproduced from [Hubel and Wiesel, 1959].

Hubel & Wiesel suggested that such increasingly complex and invariant object representations could be progressively built by integrating convergent inputs from lower levels. For instance, as illustrated in Fig. 1-4 (reproduced from [Hubel and Wiesel, 1959]), position invariance at the complex cells level, could be obtained by pooling over simple cells at the same preferred orientation but at slightly different positions.

Computer vision systems: Motivated by earlier work on perceptrons [Rosenblatt, 1962], Fukushima developed a computer vision system based on the Hubel & Wiesel model. After a series of extension [Fukushima, 1975, 1980], the network was shown to perform well in digit recognition applications [Fukushima et al., 1983]. In turn, after the development of a rigorous mathematical framework to train multi-layer network architectures with the back-propagation algorithm [Parker, 1986; Rumelhart et al., 1986; LeCun, 1986], several systems (called convolutional networks) were subsequently developed. In particular, the system developed by LeCun and colleagues was shown to perform extremely well in the

domain of digits recognition [LeCun et al., 1989, 1998] and more recently in the domain of generic object recognition [LeCun et al., 2004], face identification [Chopra et al., 2005] and for controlling an autonomous off-road vehicle [LeCun et al., 2005].

Before closing this review on biologically-inspired computer vision systems, let us briefly mention two (non-exclusive) classes of computer vision systems that are (roughly) inspired by biology and could therefore provide a plausibility proof for certain computational principles. Approaches that rely on qualitative image-based representations, *e.g.*, ordinal encoding, constitute one such type. Indeed Thorpe and colleagues have argued for some time that such representation (based on temporal order coding, see Section C.2) could be used in visual cortex. As a plausibility proof they designed a very fast computer vision system called *SpikeNet* [Thorpe and Gautrais, 1997; VanRullen et al., 1998; Gautrais and Thorpe, 1998; Delorme and Thorpe, 2001; Thorpe et al., 2001a; Thorpe, 2002]. Qualitative encoding schemes have been shown to be particularly robust to image degradations such as changes in light and illumination, for stereo matching [Bhat and Nayar, 1998], object recognition [Sali and Ullman, 1999; Sinha, 2002], iris identification [Sun et al., 2004]. Additionally [Sadr et al., 2002] showed that image reconstruction was possible based on ordinal representations.

Other computer vision systems related to biology are the *component-based* or also called *part-based* systems, see [Mohan et al., 2001; Heisele et al., 2001c; Ullman et al., 2002; Torralba and Oliva, 2003] and also [Lowe, 2000]. Those hierarchical systems contain two layers: In the first layer, the outputs of a few component-detectors (*e.g.*, eye-, nose-, mouth-detectors in the case of face detection) are locally maximized and further passed to the second layer that performs the final verification. While such systems are only vaguely mimicking the visual cortex and lack a direct implementation in terms of plausible neural mechanisms, they may however provide insights and design principles for biological vision [Serre et al., 2004a; Ullman et al., 2002].

For instance, Radial Basis Function (RBF) networks [Poggio and Girosi, 1990; Poggio and Smale, 2003] are among some of the best learning algorithms, yet simple enough to be implemented with biologically plausible circuits. RBF networks combine the activity of units that are broadly tuned to one of the training examples and have been shown to generalize well to new *unseen* examples by interpolating among the *learned* examples. Poggio & Edelman have demonstrated that such network can perform view-invariant recognition

of 3D objects from just a few 2D views. This, in turn, motivated the experimental work by Logothetis *et al.* who trained monkeys to finely discriminate 3D objects (paperclips). They found a large proportion of cells in IT that were tuned to particular views of the paperclip objects presented during training as well as a small number of view-invariant cells (as suggested by [Poggio and Edelman, 1990]). The scheme was later extended to deal with time sequences for the recognition of biological motion [Giese and Poggio, 2003]. Recently Poggio & Bizzi emphasized that neurons with a bell-shaped tuning are common in cortex and suggested that the same principles could apply to different modalities (*e.g.*, motor cortex, see [Poggio and Bizzi, 2004]).

Neurobiological models: As discussed earlier, we limit our review to a special class of models of object recognition in cortex which appears to be compatible with most of the physiology and anatomy of the ventral stream of visual cortex [Riesenhuber and Poggio, 1999a]. These models share the basic idea that the visual system is a feedforward processing hierarchy where invariance ranges and complexity of preferred features grow as one ascends through the levels.

Perhaps the first outline of a neurobiological model of shape processing in the ventral stream is the model by Perrett & Oram [Perrett and Oram, 1993; Oram and Perrett, 1994] illustrated in Fig. 1-2 (see also [Gochin, 1994]). Extending Fukushima's translation-invariant *Neocognitron* [Fukushima, 1980], they showed how a generalization of the pooling mechanisms used in the Neocognitron for invariance to translation could provide scale invariance as well.

Based on the same principles, another model implementation, *VisNet*, was proposed [Wallis and Rolls, 1997]. The model extended an earlier conceptual proposal [Rolls et al., 1992] (similar to [Perrett and Oram, 1993]) and earlier implementations [Wallis et al., 1993; Rolls, 1995]. The model relied on a trace learning rule [Földiák, 1991] to learn invariances. The algorithm exploits the temporal continuity between views of a target object during the presentation of a sequence of the object undergoing a transformation. The scheme was shown to enable the network to learn translation, scale and view-invariant representations at the level of IT. Various derivations of the original trace learning rule have been proposed since [Stringer and Rolls, 2000; Rolls and Milward, 2000; Stringer and Rolls, 2002; Elliffe et al., 2002]. The most recent extension [Deco and Rolls, 2004] includes a model of the

dorsal stream to account for top-down and attentional effects (see also [Amit and Mascaro, 2003] for a model of the ventro-dorsal interaction).

Summarizing and integrating previous approaches, [Riesenhuber and Poggio, 1999a] showed that a feedforward theory could duplicate *quantitatively* the tuning and invariance properties of the so-called view-tuned cells in AIT [Logothetis et al., 1995] (see also Section B.1). The model relied on a non-linear MAX-like pooling operation as a key mechanism to provide invariance to image degradations while avoiding the *superposition* problem (*i.e.*, the simultaneous presentation of multiple weak stimuli being as strong as the activity of the preferred stimulus. Further extensions of the original model was shown to perform well on a face detection task [Serre et al., 2002] as well as a generic object recognition task [Wersing and Koerner, 2003]. The architecture of the system by Wersing & Körner is now partially designed by *evolution principals* through genetic algorithms [Schneider et al., 2005].

Von der Malsburg formulated the main criticism to this class of feedforward models also called the *feature-binding* problem [von der Malsburg, 1981, 1995, 1999]. He suggested that models that rely on spatially invariant feature-detectors, because of the lack of relative position and size information between detectors, may fail to discriminate between object composed of the same basic dictionary of features and disentangle between features from one object and features from another object or clutter (leading to potential hallucinations) [von der Malsburg, 1995]. Based on model simulations, there is now growing evidence that the binding problem is indeed not a problem and that the claim was erroneous [Riesenhuber and Poggio, 1999b; Mel and Fiser, 2000; Stringer and Rolls, 2000]. As discussed in [Riesenhuber and Poggio, 2000] a gradual and parallel increase in both the complexity of the preferred stimulus and the invariance properties of the neurons prevent to avoid an explosion in the number of units to encode a large number of objects while circumventing the binding problem.

Summary: On the one hand, there are computer vision systems that are inspired by biology and that exhibit good performance on real-world recognition problems. Yet because they lack a direct correspondence with cortical stages, such systems cannot be used to make predictions for physiologists. Alternatively, there are neurobiological models, constrained by the anatomy of the visual cortex. Yet, for the most part, these models have

only been tested on artificial simple object images (*e.g.*, paperclips presented on a blank background [Riesenhuber and Poggio, 1999a], bars [Stringer and Rolls, 2002; Elliffe et al., 2002] or letters of the alphabet in [Deco and Rolls, 2004]). When trained on natural images (*e.g.*, [Wallis and Rolls, 1997]), datasets tend to be small, images are preprocessed and the performance of such systems is never evaluated on novel unseen examples. So far, none of the neurobiologically plausible models have been tested for their recognition capabilities on large scale, real-world, image databases where objects (faces, cars, pedestrians, *etc*) undergo drastic changes in appearance and are presented on complex clutter. In particular, it remains unclear whether such architectures could explain the high level of performance achieved by the primate visual system during rapid categorization tasks [Thorpe et al., 1996].

C.2 Cortical Circuits and Key Computations

From single neurons to computational modules: As discussed earlier, given the immediate selectivity and tuning of cells (*i.e.*, within very small temporal windows of $10-30\,ms$), the underlying neural circuits have to operate on only very few spikes. This suggests that, during such temporal windows, neurons can only transmit a few bits. Yet neural networks and models of object recognition in cortex have typically relied on the *neural activity* being an analog (continuous) value. Whether or not feedforward neural networks can indeed transmit rate codes is still under debate [van Rossum et al., 2002; Litvak et al., 2003]. One way to cope with the problem of insufficient dynamic range, which does not involve firing rates, is to consider *computational modules*[2], *i.e.*, groups of n *equivalent* cells [Földiák and Young, 1995; Perrett et al., 1998; Shadlen and Newsome, 1998; Keysers et al., 2001; Serre et al., 2005a], as the basic unit of processing rather than individual neurons. The information transmitted by one stage to the next is not about how much one neuron fires but rather how many neurons of a particular type fire within temporal windows of about $10-30\,ms$ [Thorpe and Imbert, 1989; Thorpe and Fabre-Thorpe, 2001; Keysers et al., 2001; Rolls, 2004].

While the solution seems very suboptimal (having n units that encode each of the possible feature dimension at each location in the visual field), redundancy in cortical organization is a well documented fact, *e.g.*, $\approx 80-100$ neurons in a general column [Mountcastle, 1957, 1997] and even 2.5 times these numbers in V1 [Mountcastle, 1997]. Interestingly [Shadlen and Newsome, 1998] estimated that ensembles of $\approx 50-100$ neurons were suf-

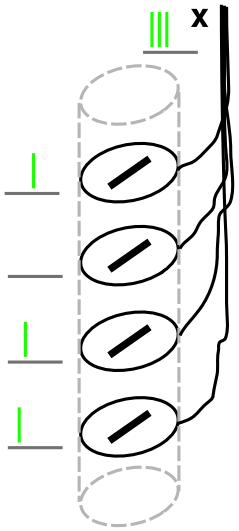


Figure 1-5: Computational modules in cortex: The basic processing unit in models of object recognition in cortex (*e.g.*, the nodes of Fig. 2-1) may correspond to *computational modules* in cortex rather than single neurons. Modules would be composed of *equivalent* cells with identical parameters and identical inputs from other units in the circuit. In addition, each cell receives an individual bias term (normally distributed background noise). Instead of $1-3$ spikes available to estimate firing rates (within the $10-30\,ms$ time window available), the postsynaptic cell now receives up to $2n$ spikes from the n neurons in the module. Such modules may correspond to *cortical columns* [Mountcastle, 1957, 1997] (or part of it). For instance because of geometric constraints, the axons of the neurons within a cortical column, are likely to contact the dendrite of a postsynaptic neuron in the same vicinity and may thus correspond to a single compartment (averaging out the activity of the module).

ficient to reliably transmit "firing rates". To paraphrase Mountcastle [Mountcastle, 1997]: "the effective unit of operation in such a distributed system is not the single neuron and its axon, but groups of cells with similar functional properties and anatomical connections". Fig. 1-5 illustrates how the dynamic range of the module is increased by a factor n compared to single neurons.

It is important to point out that the number of cells n in the module probably decreases along the visual hierarchy from V1 to IT. In early stages, a large dynamic range of the inputs is needed, whereas at the other extreme in IT, only the binary presence or absence of each critical feature has to be conveyed. A cortical column in V1 contains 2.5 times more neurons in V1 than in a column in extrastriate cortex [Mountcastle, 1997]. The basal dendritic arbors of layer III pyramidal neurons tend to become larger and more spinous towards higher cortical areas [Elston, 2003]. Contrast invariance data also provide some indirect support to the idea that the number of units in each module decreases along the hierarchy. For instance [Sclar et al., 1990] showed that the steepness of the contrast-response functions of neurons increases from LGN through V1, V2 to MT and that "cells become, in the contrast domain, progressively more like switches, being either on or off" [Lennie, 1998].

Key computations: The key computational issue in object recognition is the specificity-invariance trade-off: recognition must be able to finely discriminate between different objects or object classes while at the same time be tolerant to object transformations such as

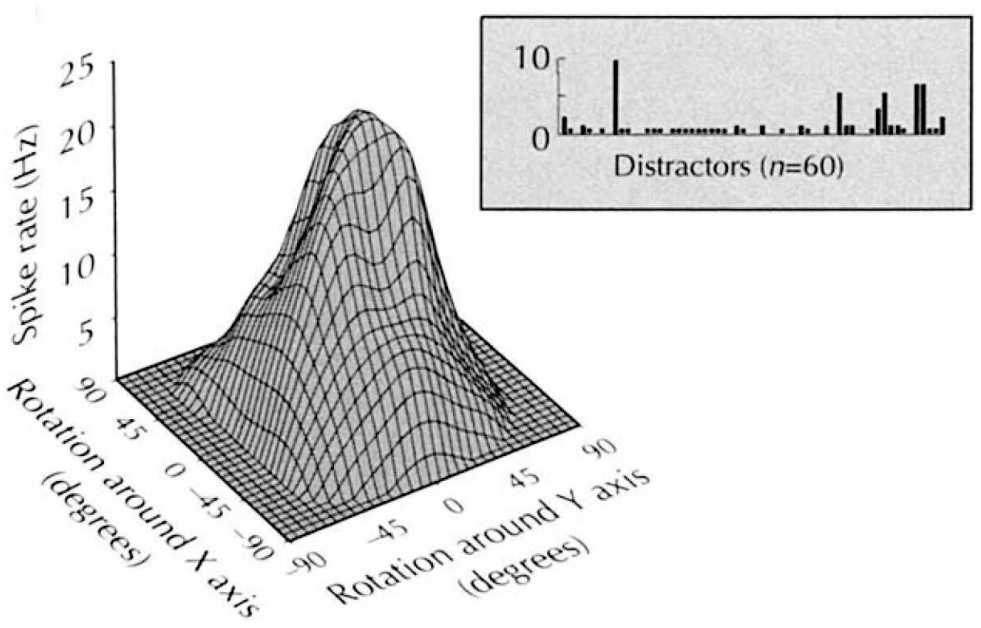


Figure 1-6: A typical bell-shaped TUNING from one cell in AIT. Illustrated here is the tuning of a particular cell to a specific view of a paperclip presented during training [Logothetis et al., 1995]. As the stimulus presented rotates away from the tuned view (along either of the two axes), the response of the cell decreases with a (Gaussian-like) bell-shaped curve. Neurons with such tuning are prevalent across cortex. [Poggio and Bizzi, 2004] argued that this may be a key feature of the *generalization* ability of cortex (*i.e.*, the ability to generalize to new unseen examples by opposition to a look-up table), see text. The figure is modified from [Logothetis et al., 1995].

scaling, translation, illumination, changes in viewpoint, changes in context and clutter, as well as non-rigid transformations (such as a change of facial expression) and, for the case of categorization, also to variations in shape within a class. Thus the main computational difficulty of object recognition is achieving a trade-off between selectivity and invariance. Theoretical considerations [Riesenhuber and Poggio, 1999a] suggested that only two functional classed of units may be necessary to achieve this trade-off:

- The *simple S* units perform a TUNING operation over their afferents to build object-selectivity. The simple *S* units receive convergent inputs from retinotopically organized units tuned to *different preferred stimuli* and combine these *subunits* with a bell-shaped tuning function, thus increasing object selectivity and complexity of the preferred stimulus.

  The analog of the TUNING in computer vision is the *template matching* operation between an input image and a stored representation. As discussed in [Poggio and Bizzi, 2004] neurons with a Gaussian-like bell-shape tuning are prevalent across cor-

tex. Eq. 1.2 offers a phenomenological model of the bell-shaped tuning found across cortex. For instance simple cells in V1 exhibit a Gaussian tuning around their preferred orientation (see Chapter 2) or as pointed out earlier in Section B cells in AIT are typically tuned around a particular view of their preferred object. Fig. 1-6 illustrates the bell-shape tuning of a typical AIT cell from [Logothetis et al., 1995] to a particular view of a paperclip presented during training. From the computational point of view, Gaussian-like tuning profiles may be key in the generalization ability of cortex. Networks that combine the activity of several units tuned with a Gaussian profile to different training examples have proved to be powerful learning scheme [Poggio and Girosi, 1990; Poggio and Smale, 2003].

- The *complex C* units perform a MAX-like [3] operation over their afferents to gain invariance to several object transformations. The complex *C* units receive convergent inputs from retinotopically organized *S* units tuned to the *same preferred stimuli* but at slightly different positions and scales with a MAX-like operation, thereby introducing tolerance to scale and translation. Fig. 1-7 shows an example of a MAX operation being performed at the level of a V1 complex cell.

  MAX functions are commonly used in signal processing (*e.g.*, selecting peak correlations) to filter noise out. The existence of a MAX operation in visual cortex was predicted by [Riesenhuber and Poggio, 1999a] from theoretical arguments (and limited experimental evidence [Sato, 1989] and was later supported experimentally in V4 [Gawne and Martin, 2002] and in V1 at the complex cell level [Lampl et al., 2004]. Fig. 1-7 (reproduced from [Lampl et al., 2004] illustrates how a complex cell may combine the response of oriented retinotopically organized subunits (presumably simple cells) at the same preferred orientation with a MAX pooling mechanism.

As discussed earlier, a *gradual* increase in both selectivity and scale (as observed along the ventral stream) is critical to avoid both a combinatorial explosion in the number of units, and the binding problem between features. Below we shortly give idealized mathematical approximations of the operations and discussed possible cortical circuits to implement them.

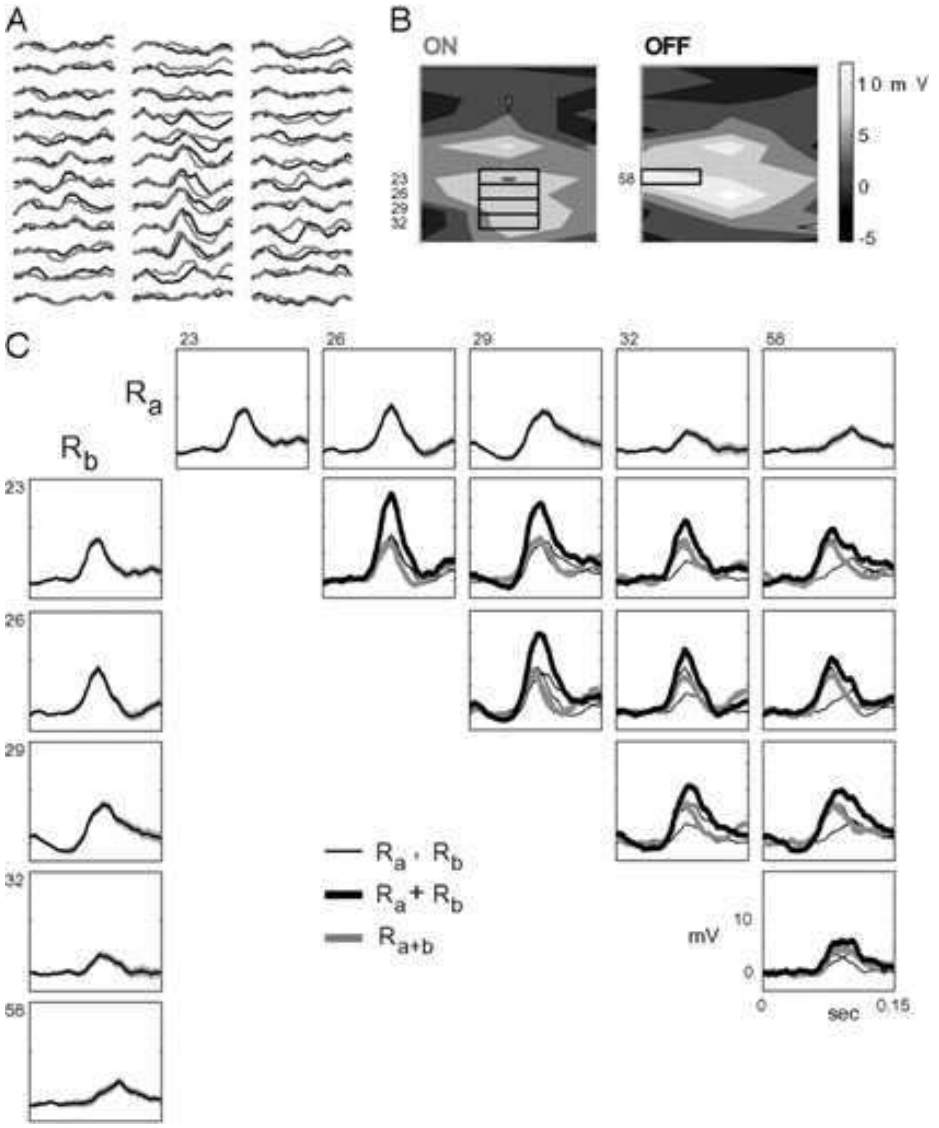


Figure 1-7: MAX operation from a complex cell in area 17 of the cat. Illustrated is the response of a complex cell to the simultaneous presentation of two bars (see [Lampl et al., 2004]) for details). A: average membrane potential measured from the response of the cell to bars of the optimal orientation. Black traces are the responses to dark bars (OFF responses) and gray traces are the responses to bright bars (ON responses). B: intensity plots obtained from the mean potentials. C: cell responses to each of the selected bars shown in B by thick lines around the rectangles. Lines in the 1^{st} row and 1^{st} column panels are the averaged responses to the presentation of a single bar, and the shaded area shows the mean ($\pm SE$). The inner panels present the response of the cell to the simultaneous presentation of the 2 bars whose positions are given by the corresponding column and row (gray traces), the responses to the 2 stimuli presented individually (thin black traces) and the linear sum of the 2 individual responses (thick black traces). Modified from [Lampl et al., 2004]

Idealized mathematical descriptions of the two operations: In the following, we denote y the response of a unit (simple or complex). The set of inputs to the cell (*i.e.*, presynaptic units) are denoted with subscripts $j = 1 \ldots N \in \mathcal{N}$. When presented with a pattern of activity $\mathbf{x} = (x_1, \ldots x_N)$ as input, an idealized – and static – description of the unit response y is given by:

$$y = \max_{j \in \mathcal{N}} x_j \tag{1.1}$$

As mentioned earlier, for a complex cell, The inputs x_j to the units are retinotopically organized (selected from an $m \times m$ grid of afferents with the same selectivity). For instance in the case of a V1-like complex cell tuned to an horizontal bar, all subunits are tuned to an horizontal bar but at slightly different positions and spatial frequency (or equivalently scale or bar dimension). Similarly, an idealized description of a simple unit response is given by:

$$y = \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(w_j - x_j)^2\right) \tag{1.2}$$

σ defines the sharpness of the TUNING of the unit around its preferred stimulus (also called *center* for RBF networks [Poggio and Girosi, 1990]) corresponding to the synaptic strengths $\mathbf{w} = (w_1, \ldots w_n)$. As for complex cells, the subunits of the simple cells are also retinotopically organized (selected from an $m \times m$ grid of possible afferents). But, in contrast with complex cells, the subunits of a simple cell can be with different selectivities to increase the complexity of the preferred stimulus. For instance, for S_2 units the subunits are V1-like complex cells (with a small range of invariance to position and scale) at different preferred orientations. Eq. 1.2 accounts for the bell-shaped tuning of cells found across cortex [Poggio and Bizzi, 2004]. That is, the response of the unit is maximal ($y = 1$) when the current pattern of input $\mathbf{x}$ matches exactly the synaptic weights $\mathbf{w}$ (for instance the frontal view of a face) and decreases with a bell-shaped profile as the pattern of input becomes more dissimilar (as the face is rotated away from the profile view).

Both of those mathematical descriptions are only meant to describe the response behavior of cells at a phenomenological level. [Yu et al., 2002] described several circuits that

could compute an approximation of a MAX called a SOFTMAX. [Kouh and Poggio, 2004] investigated possible approximations to Gaussian functions and found that, in high dimensional space, a Gaussian function can be well approximated by a normalized dot-product passed through a sigmoid [Kouh and Poggio, 2004; Maruyama et al., 1991, 1992]. Mathematically, a normalized dot-product and a softmax, take essentially the same general form, that is:

$$y = \frac{\sum_{j=1}^{n} w_j^* \, x_j^p}{k + \left(\sum_{j=1}^{n} x_j^q \right)^r}, \tag{1.3}$$

where $k << 1$ is a constant to avoid zero-divisions and p, q and r represent the static non-linearities in the underlying neural circuit. Such nonlinearity may correspond to different regimes on the $f - I$ curve of the presynaptic neurons such that different operating ranges provide different degrees of nonlinearities (from near-linearity to steep non-linearity). An extra sigmoid transfer function on the output $g(y) = 1/(1 + \exp^{\alpha(y-\beta)})$ controls the sharpness of the unit response. By adjusting these non-linearities, Eq. 1.3 can approximate better a MAX or a TUNING function:

- When $\mathbf{p} \lessapprox \mathbf{qr}$, the unit approximates a Gaussian-like TUNING, *i.e.*, its response y will have a peak around some value proportional to the input vector $\mathbf{w} = (w_1, \ldots, w_N)$. For instance, when $p = 1$, $q = 2$ and $r = 1/2$, the circuits perform a normalized dot-product with an L_2 norm, which with the addition of a bias term may approximate a Gaussian function very closely (see [Kouh and Poggio, 2004; Serre et al., 2005a] for details). Indeed when all vectors are normalized, *i.e.*, $||\mathbf{x}||^2 = ||\mathbf{w}||^2 = 1$, the approximation is exact and for any $\mathbf{w}$, one can compute $\mathbf{w}^*$ such that Eq. 1.2 and Eq. 1.3 are strictly equivalent (see [Maruyama et al., 1991, 1992]).

- When $\mathbf{p} \gtrapprox \mathbf{q} + \mathbf{1}$ ($\mathbf{w_j} \approx \mathbf{1}$), the unit approximate a MAX function very closely for larger q values (see [Yu et al., 2002], the quality of the approximation also increases as the inputs become more dissimilar). For instance, $r \approx 1$, $p \approx 1$, $q \approx 2$ gives a good approximation of the MAX (see [Serre et al., 2005a] for details).

Biophysical considerations: The fact that both the MAX and TUNING functions can be described by the same equation strongly suggests that they may be implemented by the same biophysical mechanisms. Indeed by simply rearranging the terms in Eq. 1.3 a possible circuitry becomes more apparent:

$$y = \sum_{j=1}^{n} w_j^* \frac{x_j^p}{k + \left(\sum_{j=1}^{n} x_j^q\right)^r} = \sum_{j=1}^{n} w_j^* \frac{x_j^p}{Pool}, \tag{1.4}$$

The equation above suggests that the operation could be carried out by a *divisive normalization* followed by *weighted sum*. Normalization mechanisms (also commonly referred to as *gain control*) in this case, can be achieved by a feedforward (or recurrent) shunting inhibition [Torre and Poggio, 1978; Reichardt et al., 1983; Carandini and Heeger, 1994]. For the past two decades several studies (in V1 for the most part) have provided evidence for the involvement of GABAergic circuits in shaping the response of neurons [Sillito, 1984; Douglas and Martin, 1991; Ferster and Miller, 2000]. Direct evidence for the existence of divisive inhibition comes from an intracellular recording study in V1 [Borg-Graham and Fregnac, 1998]. [Wilson et al., 94] also showed the existence of neighboring pairs of pyramidal cells / fast-spiking interneurons (presumably inhibitory) in the prefrontal cortex with inverted responses (*i.e.*, phased excitatory/inhibitory responses). The pyramidal cell could provide the substrate for the weighted sum while the fast-spiking neuron would provide the normalization term.

Plausible biophysical circuits based on feedforward or feedback shunting inhibition were proposed that could implement Eq. 1.3 [Yu et al., 2002; Serre et al., 2005a]. A possible (feedforward) circuit is sketched in Fig. 1-8 (reproduced from [Serre et al., 2005a]). The nodes x_1 and x_2 each represent a computational module composed of n *equivalent* units, *i.e.*, units with identical parameters that share the same afferents, but in addition each of the unit in the module receives an extra normally distributed background input (see Fig. 1-5). A detailed implementation of the circuit using parameters from experimental data [Destexhe et al., 1998] was shown to approximate well the two operations (with different parameter values) over a range of input values (see [Serre et al., 2005a] for details).

Thorpe and colleagues described another related proposal based on the timing of the arrival of spikes from a group of neurons [Thorpe and Gautrais, 1997; Thorpe et al., 2001a]

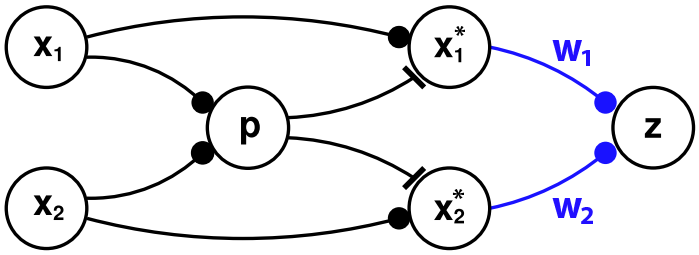


Figure 1-8: A possible cortical circuit for TUNING and MAX operations proposed by Knoblich & Poggio. The nodes in the circuit correspond to computational modules composed of equivalent units, see Fig. 1-5. Preliminary results suggest that depending on the balance between excitation and inhibition (see discussion on Eq. 1.3), the circuit can approximate a MAX or a TUNING operation (see [Serre et al., 2005a] for details).

(see [Rousselet et al., 2004a] for a recent review). The circuit proposed is very similar to the one in Fig. 1-8. The main difference is that, in the model by Thorpe, the basic element is a single neuron (*i.e.*, the model relies on individual spikes) whereas in Fig. 1-8 it is a module of n identical neurons (*i.e.*, the model relies on the firing rate produced by the ensemble of neurons). There is now limited evidence for such type of encoding in part from the somatosensory system [Johansson and Birznieks, 2004] (see [VanRullen et al., 2005] for a review).

D Original Contributions

D.1 Learning a Dictionary of Shape-Components in Visual Cortex

The model described in Chapter 2 builds upon several existing neurobiological models and conceptual proposals (see Section C) and, in particular, extends significantly an earlier approach by [Riesenhuber and Poggio, 1999a]. One of the key new aspect of the model is the learning of a generic dictionary of shape-components from V2 to IT, which provides a rich representation to task-specific categorization circuits in higher brain areas. Importantly, the hierarchical architecture builds progressively more invariance to position and scale while preserving the selectivity of the units. This vocabulary of tuned units is learned from natural images during a developmental-like, unsupervised learning stage in which each unit in the intermediate layers becomes tuned to a different patch of a natural image.

The model is characterized by a large number of tuned units across the hierarchical architecture of the model which are learned from natural images and represent a redundant

dictionary of fragment-like features that span a range of selectivities and invariances. As a result of this new learning stage, the new architecture contains a total of ~ 10 million tuned units. At the top, the classification units rely on a dictionary of $\sim 6,000$ units tuned to image features with different levels of selectivities and invariances. This is $2-3$ orders of magnitude larger than the number of features used by both biological models as well as state-of-the-art computer vision systems that typically rely on 10-100 features.

D.2 Comparison with Neural Data

As described in Chapter 3, one major advance is that the proposed model is significantly closer to the anatomy and the physiology of visual cortex with more layers (reflecting PIT as well as V4) and with a looser hierarchy (reflecting the bypass connections from V2 to PIT and V4 to AIT [Nakamura et al., 1993]). In particular we show in Chapter 3 that model units are qualitatively and quantitatively consistent with several properties of cells in V1, V4, and IT. The most significant result is that the tuning of the units in intermediate stages of the model that are learned from natural images agrees with data from V4 [Reynolds et al., 1999] about the response of neurons to combinations of simple two-bar stimuli (within the receptive field of the S_2 units). Some of the C_2 units in the model show a tuning for boundary conformations [Pasupathy and Connor, 2001] which is consistent with recordings from V4 (Serre, Cadieu, Kouh and Poggio, in prep). In addition, unlike the original model [Riesenhuber and Poggio, 1999a], all the V1 parameters are derived exclusively from available V1 data and do not depend – as they did in part in the original HMAX model – from the requirement of fitting the benchmark paperclip recognition experiments.

D.3 Comparison with Computer Vision Systems

As described in Chapter 4, another major advance achieved by the model is that, not only does the proposed architecture duplicates the tuning properties of neurons in various brain areas when probed with artificial stimuli, but, it can also handle the recognition of objects in the real-world, to the extent of competing with the best computer vision systems [Serre et al., 2005b, 2006b]. We also show that a generic dictionary of shape-tuned units learned from a set of natural images unrelated to any categorization task can support the recogni-

tion of many different object categories. In addition, we show that the model is remarkably robust to parameter values, detailed wiring and even exact form of the two basic operations and of the learning rule.

D.4 Comparison with Human Observers

The most significant result is described in Chapter 5. We compare the performance of the model and the performance of human observers in a rapid animal *vs.* non-animal recognition task for which recognition is fast and cortical back-projections are likely to be inactive. Results indicate that the model predicts human performance extremely well when the delay between the stimulus and the mask is about $50\ ms$ (Serre, Oliva & Poggio, in prep). This suggests that cortical back-projections may not play a significant role when the time interval is in this range, and the model may therefore provide a satisfactory description of the feedforward path.

Notes

[1] `http://bluebrainproject.epfl.ch`

[2] In [Serre et al., 2005a], we used the term *cable* instead of computational module, which consists of several *wires* (single axons).

[3] The MAX-like operation does not need to be exact. Indeed, preliminary results suggest that an *average* pooling mechanism may still provide a scale and translation invariant representation at the level of IT with minimal loss in recognition performance.

[4] Note that a more general form of normalization in Eq. 1.3 would involve another set of synaptic weights $\tilde{\mathbf{w}}$ in the denominator, as explored in a few different contexts such as to increase the independence of correlated signals [Heeger et al., 1996; Schwartz and Simoncelli, 2001] and the biased competition model of attention [Reynolds et al., 1999].

[5] An alternative to the tuning operation based on the *sigmoid of a normalized dot-product* (see Eq. 1.3) is a *sigmoid of a dot-product* that is:

$$y = g\left(\sum_{j=1}^{n} w_j \, x_j^p\right), \tag{1.5}$$

where g is a sigmoid function given (see above). Eq. 1.5 is less flexible than Eq. 1.3 which may provide tuning to any arbitrary pattern of activations irrespective of the overall magnitudes of the input activations. On the other hand, the dot-product tuning does not require any inhibitory elements and may thus be simpler to build. Also, with a very large number of inputs (high dimensional tuning), the total activation of the normalization pool, or the denominator in Eq. 1.3, would be more or less constant for different input patterns, and hence, the dot product and the normalized dot-product may behave very similarly. In other words, the normalization operation may only be necessary to build a robust tuning behavior with a small number of inputs, *e.g.*, in early stages such as V1. It is conceivable that both Eq. 1.3 and Eq. 1.5 are used for tuning, with Eq. 1.5 more likely in later stages of the visual pathway. In Chapter 4, we confirm that the model exhibits qualitatively similar results across several categorization tasks with either operations at the level of the top layers.

Chapter 2

Theory and Basic Model Implementation

In Chapter 1, we previously described a core of knowledge, accumulated over the past 40 years, about the organization and architecture of the ventral stream of visual cortex. In this Chapter, we describe a theory which accounts for these basic facts and a model implementation derived hereafter which is faithful to the anatomy and physiology of the ventral stream of visual cortex. Consistent with [Riesenhuber and Poggio, 2000], the model is composed of two key components: First a generic dictionary of shape components is extracted from V1 to IT and provide a translation and scale invariant representation that can be used by higher areas such as PFC to train and maintain task-specific circuits for the recognition of different object categories.

Section A describes how this dictionary of shape components map into circuits and cortical areas of the primate visual cortex. In Section B we describe how this vocabulary of shape-tuned units can be learned, in a development-like unsupervised way, from natural images. In Section C we suggest how task-specific circuits can be built upon this representation. Finally, in Section D, we discuss important aspects of the model, including a computational analysis as well as possible connections with machine learning and computer vision.

A Building a Dictionary of Shape-Components from V1 to IT

The theory we propose significantly extends an earlier model by [Riesenhuber and Poggio, 1999a] and builds upon several conceptual proposals [Hubel and Wiesel, 1959; Perrett and Oram, 1993; Rolls, 1995; Hochstein and Ahissar, 2002], computer vision systems [LeCun et al., 1989; Fukushima, 1980; Mel, 1997; Thorpe, 2002; Ullman et al., 2002; Wersing and Koerner, 2003; LeCun et al., 2004] and models of object recognition in cortex [Wallis and Rolls, 1997; Amit and Mascaro, 2003].

A model implementation that reflects the general organization of the ventral stream of visual cortex from V1 and V2 through V4, TE, TEO and PFC is sketched in Fig. 2-1. The (tentative) correspondence between the functional primitives and various stages of the model (right) and cortical stages in the primate visual system (left, modified from Van Essen & Ungerleider [Gross, 1998]) is color coded. For instance, the S_1 and C_1 layers are filled in red as their cortical homologues (areas V1/V2). Along the hierarchy, from V1 to IT, two functional stages are interleaved to provide a basic object representation in IT that be read out by higher cortical stages to perform a large array of visual tasks. Those two stages are:

- Various stages of *simple* (S) units (plain circles and arrows) build an increasingly complex and specific representation by combining the response of several subunits with different selectivities with a TUNING operation (see Eq. 1.2 and Chapter 1 for details);

- Various stages of *complex* (C) units (dashed circles and arrows) build an increasingly invariant representation (to position and scale) by combining the response of several subunits with the same selectivity but at slightly different position and scales with a MAX-like operation (see Eq. 1.1 and Chapter 1 for details).

In the following, starting with V1, we provide a thorough description of the different stages of the model.

A.1 Simple and Complex Cells in V1

The input to the model is a gray-value image. Typically, images used range between 140×140 pixels and 256×256 pixels corresponding to about 4^o and 7^o of visual angle respectively.[1]

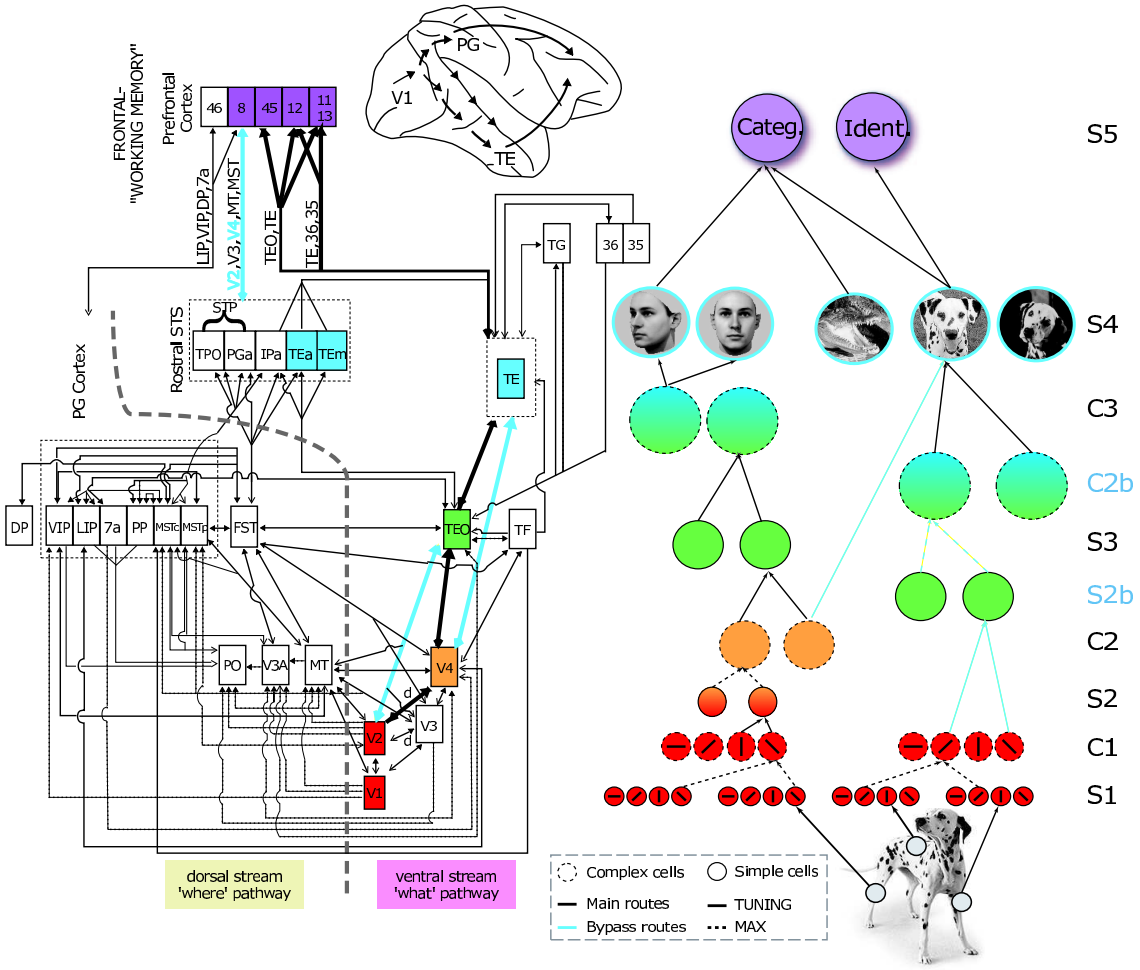


Figure 2-1: Tentative mapping between (right) the functional primitives and layers of the model and (left) cortical stages in the primate visual system (modified from Van Essen & Ungerleider [Gross, 1998]). The correspondences are illustrated with colors (see text). Stages of *simple* (*S*) units (plain circles) build an increasingly complex and specific representation by combining the response of several subunits with different selectivities (see text) and exhibit a Gaussian-like TUNING (see Eq. 1.2). Layers of simple units are interleaved with layers of *complex* units (dotted circles) which combine several units with similar selectivities but slightly different positions and scales to increase invariance to object transformations (pooling over scales is not shown in the figure). The pooling operation at the complex unit level is a MAX-like operation [Gawne and Martin, 2002; Lampl et al., 2004]. Both operations may be performed by the same local recurrent circuits of lateral inhibition (see Chapter 1). Black arrows correspond to the main route providing the main inputs to IT (the final purely visual cortical area in the ventral stream). Light blue arrows illustrate the bypass routes (see text). Learning at the level of the *S* units from V2 up to IT is assumed to be stimulus-driven. Visual experience shape the TUNING of the units through task-independent mechanisms (see Section B). Supervised learning occurs at the level of the task-specific circuits in PFC (two sets of possible circuits for two of the many different recognition tasks – identification and categorization – are indicated). The model which is feedforward (apart from local recurrent circuits) attempts to describe the initial stage of visual processing, *i.e.*, *immediate recognition*, corresponding to the first 150 milliseconds of visual recognition.

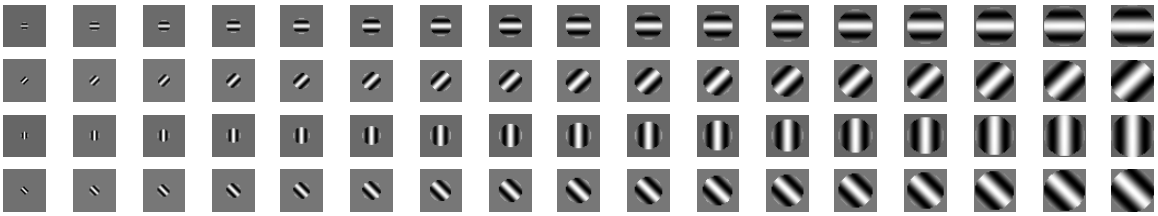


Figure 2-2: Receptive field organization of the S_1 units. There are 136 different types of S_1 units: 2 phases $\times$ 4 orientations $\times$ 17 sizes (or equivalently peak frequencies). Only units at one phase are shown but the population also includes filters of the opposite phase. Receptive field sizes range between $0.2^o - 1.1^o$ (typical values for cortex range between ($\approx 0.1^o - 1^o$, see [Schiller et al., 1976e; Hubel and Wiesel, 1965]). Peak frequencies are in the range $1.6 - 9.8$ cycles/deg.

S_1 units: The input image is first analyzed by an a multi-dimensional array of simple S_1 units which correspond to the classical V1 simple cells of Hubel & Wiesel (see Chapter 1). Model S_1 units follow the basic model of simple cells, *i.e.*, half-rectified filters consisting of aligned and alternating ON and OFF subregions, which share a common axis of elongation that defines the cell preferred orientation.[2]

The population of S_1 units consists in 96 types of units, *i.e.*, 2 phases $\times$ 4 orientations $\times$ 17 sizes (or equivalently peak spatial frequencies[3]). Fig. 2-2 shows the different weight vectors corresponding to the different types of units (only one phase shown). Each portion of the visual field (*i.e.*, each pixel location in the input image) is analyzed by a full set of the 96 unit types which may correspond to one macro-column in V1 [Hubel and Wiesel, 1977]. This is illustrated in Fig. 2-3.[4]

S_1 units, like other simple units in the model, perform a TUNING operation between the incoming pattern of input $\mathbf{x}$ and their weight vector $\mathbf{w}$. The response of a S_1 unit is maximal when $\mathbf{x}$ matches $\mathbf{w}$ exactly. Typically a high response is elicited when the orientation of the stimulus, *e.g.*, a bar of optimal width and height, an edge or a grating at the optimal spatial frequency, matches the filter orientation and the response drops-off as the orientation of the stimulus and the filter becomes more dissimilar (see Fig. 2-4).

Mathematically the weight vector $\mathbf{w}$ of the S_1 units take the form of a Gabor function [Gabor, 1946] (see Eq. A.3 in Appendix A), which have been shown to provide a good model of simple cell receptive fields [Marcelja, 1980; Daugman, 1980a,b; Hawken and Parker, 1987; Jones and Palmer, 1987]. In setting the S_1 unit parameters we tried to generate a population of units that match the bulk of parafoveal cells as closely as possible (see Chapter 3). The complete parameter set used to generate the population of S_1 units is

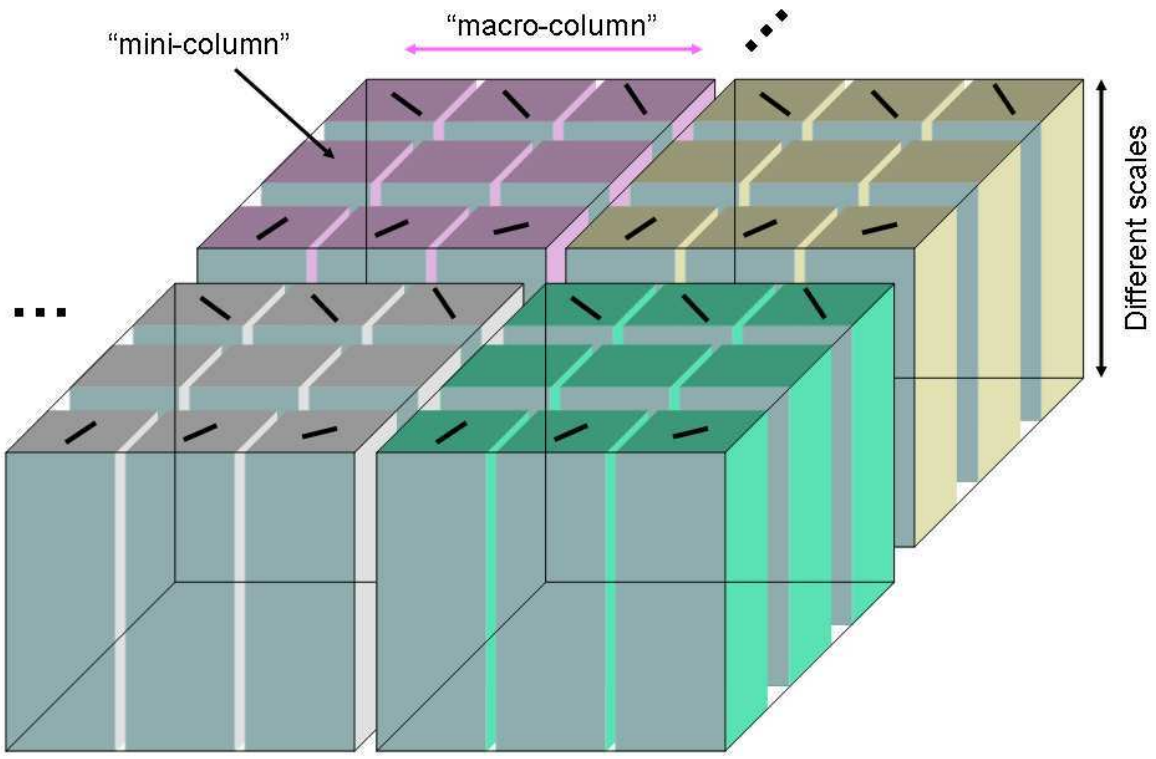


Figure 2-3: Functional "columnar" organization in the model. Each basic *mini-column* contains a set of units all with the same selectivities, *i.e.,* sharing the same weight vector $\mathbf{w}$ (*e.g.,* a bar at a particular orientation at the S_1 level) but different scales (*e.g.,* 17 different scales/peak frequencies at the S_1 level). Each portion of the visual field is analyzed by a *macro-column* which contains all types of *mini-columns* (*e.g.,* 4 different orientations and 2 phases in the S_1 case). The same organization is repeated in all layers of the model with increasingly complex and invariant units. Also note that there is a high degree of overlap in the portions of the visual field covered by neighboring macro-columns. Importantly note that we refer to columns in the model as *functional primitives* by analogy to the organization of visual cortex. Whether or not such functional columns in the model correspond to structural columns in cortex is still an open question.

given in Appendix A and a comparison between model S_1 units and V1 parafoveal cells is summarized in Chapter 3.

C_1 units: The next C_1 level corresponds to striate complex cells [Hubel and Wiesel, 1959]. Each of the complex C_1 unit receives the outputs of a group of simple S_1 units from the first layer with the same preferred orientation (and two opposite phases) but at slightly different positions and sizes (or peak frequencies). The operation by which the S_1 unit responses are combined at the C_1 level is a nonlinear MAX-like operation such that the response of the C_1 unit is determined by the strongest of all its inputs. As discussed in Chapter 1, this non-linear pooling operation provides an increase in the tolerance to changes in po-

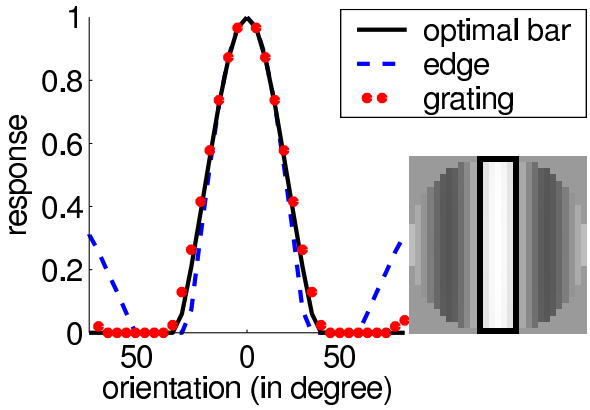


Figure 2-4: One S_1 unit and its corresponding orientation tuning curve obtained with three classical stimuli, *i.e.*, optimal bar (see black rectangle superimposed on the unit receptive field), optimal grating and edge.

sition and scale from the S_1 to the C_1 layers while avoiding the superposition problem, *e.g.*, a unit performing a SUM over its inputs could not discriminate between the presence of many weak stimuli and the presence of its preferred (optimal) stimulus.

This principle is illustrated in Fig. 2-5. For clarity we depict pooling over space and pooling over position as two separate mechanisms but in the model implementation both pooling over space and scale are performed in one single operation. By pooling over S_1 units at slightly different positions but same preferred orientation, the corresponding C_1 unit becomes insensitive to the location of the stimulus within its receptive field, which is a hallmark of the complex cells [Hubel and Wiesel, 1959, 1962, 1965, 1968]. The effect of the pooling over S_1 units at slightly different peak frequencies (or scale) is a broadening of the frequency bandwidth from S_1 to C_1 units also in agreement with physiology [Hubel and Wiesel, 1968; Schiller et al., 1976e; DeValois et al., 1982a] (see also Chapter 3), *i.e.*, the larger the pooling range, the broader the frequency bandwidth.

Similarly the size of the spatial neighborhood over which the C_1 units pool over determines its receptive field size. From S_1 to C_1 receptive field sizes double (from $0.2^o - 1.0^o$ in S_1 layer to $0.4^o - 2.0^o$ in C_1 layer) [5]. As for the S_1 units, the values of the two pooling parameters were manually adjusted so that the tuning properties of the corresponding C_1 units match closely those of V1 parafoveal complex cells (see Chapter 3). A summary of the C_1 parameter values can be found in Appendix A.

The plausibility of such MAX pooling mechanisms over simple cells at different positions at the complex cell level has been more directly tested by Lampl *et al.* [Lampl et al., 2004] via intracellular recordings from area 17 of the cat (homologous to V1 in monkey). Fig. 1-7 (reproduced from [Lampl et al., 2004]) illustrates one such cortical complex cell

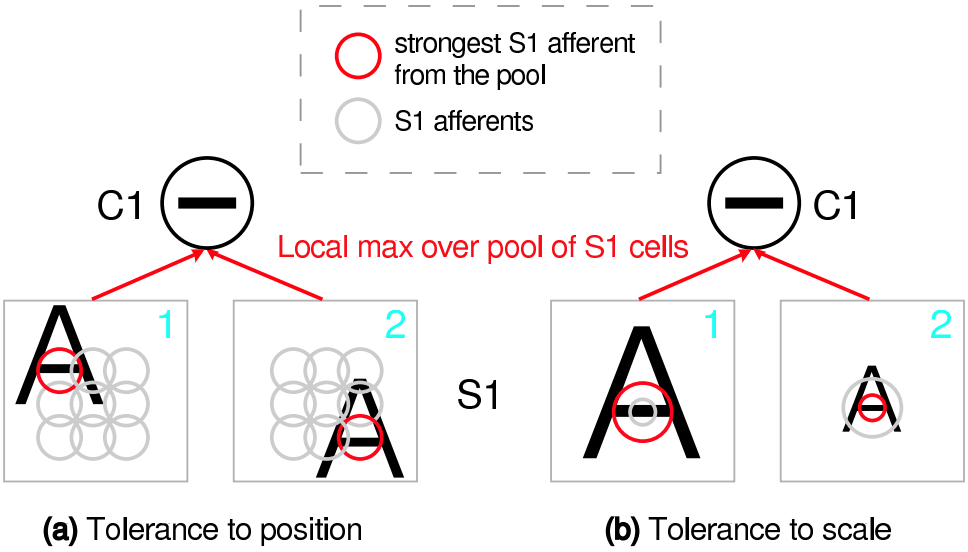


Figure 2-5: How tolerance to scale **(b)** and position **(a)** is gained from the S_1 to the C_1 layer: Each C_1 unit receives its inputs from S_1 units at the same preferred orientation (*e.g.*, 0^o) but (two) slightly different peak frequencies and positions (*e.g.*, within a small 3×3 spatial neighborhood). When the input letter is shifted from position 1 to 2 **(a)**, it activates in turn S_1 units at two different positions. By pooling the activity of all the units in the neighborhood the C_1 unit becomes insensitive to the location of the stimulus. Similarly for invariance to scale **(b)**, when the size of the letter is reduced from 1 to 2, the S_1 unit maximally activated changes from the larger to the smaller S_1 unit. By pooling the activity of S_1 units at different scales (or peak frequencies) the C_1 unit becomes insensitive to small changes in scale. For illustration purpose, we show the pooling over space and scale as separate processes but in the model implementation this is done in one stage.

which performs a MAX operation over its afferents: The response of the cell to two simultaneously presented bars is determined by the strongest response of the cell when the two bars are presented in isolation.

A.2 Beyond V1: Features of Moderate Complexity

In the next stages of the model, by interleaving these two operations, *i.e.*, MAX over retinotopically organized inputs with the same preferred stimulus but slightly different positions and scales and TUNING over inputs with different preferred stimuli, an increasingly complex and invariant representation is built [Kobatake et al., 1998]. From V1, the visual information is routed to V2, V4 and IT, which has been shown to be critical in the ability of primates to perform invariant recognition. This is done via two routes: a *main* route that follows the hierarchy of cortical stages strictly (*i.e.*, step-by-step) as well as several *bypass* routes which skip some of the stages (see Fig. 2-1). We suggest that bypass routes may help create a richer repertoire of features with various degrees of selectivities and invariances.

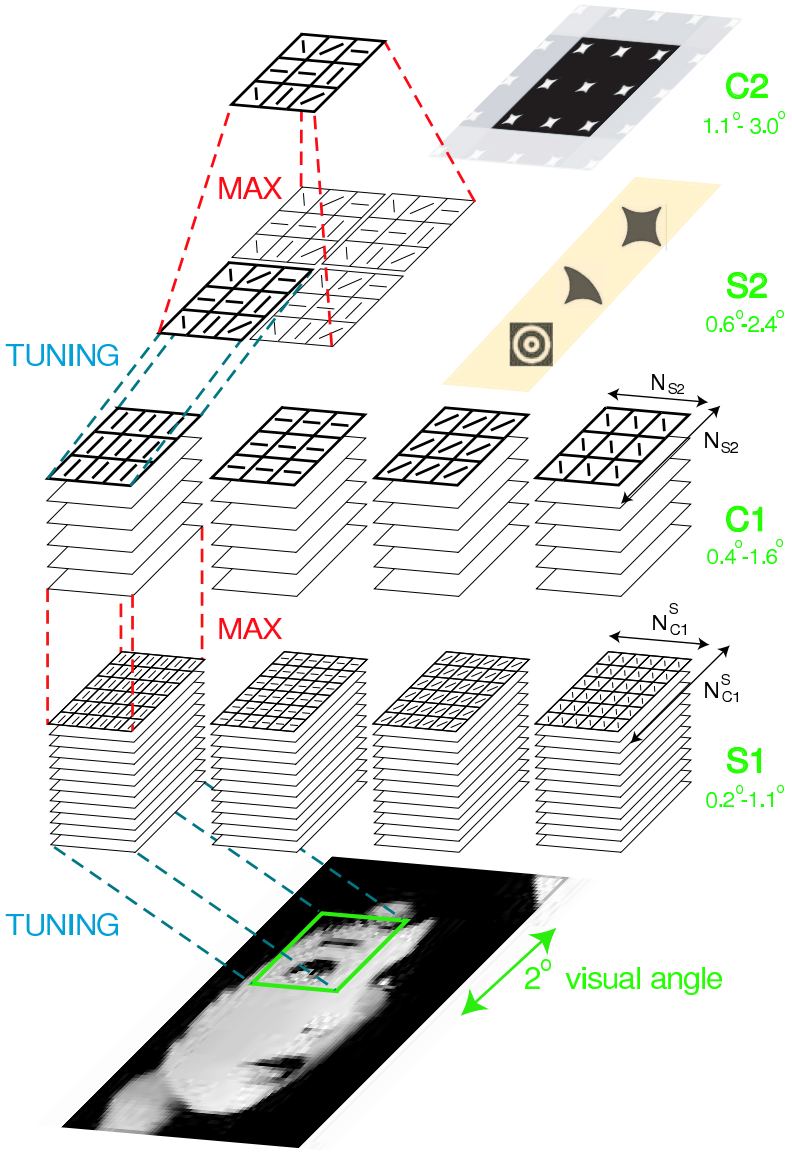


Figure 2-6: Building S_2 and C_2 units. A gray-value input image is first analyzed by functionally organized (see Fig 2-3) S_1 units at all locations. At the next C_1 layer, a local MAX pooling operation is taken over retinotopically organized S_1 units at neighboring positions and scales but with the *same* preferred orientation (presumably within adjacent macro-columns) to increase invariance to position and scale. In the next S_2 stage, a TUNING operation is taken over C_1 units at *different* preferred orientations to increase the complexity of the optimal stimulus: The S_2 receptive fields thus correspond to the nonlinear combination of V1-like oriented subunits. S_2 units are selective for features of moderate complexity [Kobatake et al., 1998] (examples shown in yellow next to the S_2 unit). We only show one type of S_2 units but in the model implementation, by considering different combinations of C_1 units (learned from natural images), we obtained $n \approx 1,000$ different types of S_2 units. Also note that S_2 units are also organized in *columns* (not shown here) such that each column contains all n types of S_2 units at different scales and analyzes a small region of the visual field. A local MAX pooling operation is performed over S_2 units with the same selectivity over neighboring positions and scales to yield the C_2 unit responses.

Main route

At the S_2 level, units pool the activities of several retinotopically organized complex C_1 units at different preferred orientations over a small neighborhood (again the size of the neighborhoods determine the size of the receptive field of the S_2 unit). The computation performed during pooling is the TUNING operation. As a result, from C_1 to S_2 units, both the selectivity of the units and the complexity of their preferred stimuli are increased.

This is illustrated in Fig. 2-6. At the C_1 level units are selective for single bars at a particular orientation, whereas at the S_2 level, units becomes selective to more complex patterns – such as the combination of oriented bars to form contours or boundary-conformations [Pasupathy and Connor, 2001] (see Chapter 3). Receptive field sizes at the S_2 level range between $0.6^o - 2.4^o$.

Beyond the S_2 layer, the tuning (*i.e.*, the input weights) of all S units is learned, in an unsupervised manner, from natural images (see Section B). In Fig. 2-6 only one type of S_2 unit is shown but in the model implementation, there is $n \approx 1,000$ types of S_2 units that correspond to different combinations of complex C_1 unit responses. Also in the model implementation, the $S2$ layer is organized in overlapping columns such that a small part of the visual field is analyzed by one such column which contains all n unit types at all scales (*i.e.*, 8 different scales coming from the 8 C_1 scales).

In the next C_2 stage, units pool over S_2 units that are tuned to the same preferred stimulus (they correspond to the same combination of C_1 units and therefore share the same weight vector $\mathbf{w}$) but at slightly different positions and scales. C_2 units are therefore selective for the same stimulus as their afferents S_2 units. Yet they are less sensitive to the position and scale of the stimulus within their receptive fields. Receptive field sizes at the C_2 level range between $1.1^o - 3.0^o$. As indicated in Fig. 2-1 and as we show in more detail in Chapter 3 (see also [Cadieu, 2005]), we found that the tuning of model C_2 units (and their invariance properties) to different standard stimuli such as Cartesian and non-Cartesian gratings, two-bar stimuli and boundary conformation stimuli is compatible with data from V4 [Gallant et al., 1996; Pasupathy and Connor, 2001; Reynolds et al., 1999].

The precise correspondence of the S_2 units (recall that S_2 units exhibit the same selectivity as the C_2 units but exhibit a lesser range of invariance) is less constrained. We speculate that S_2 units are likely to be found principally in layer IV of area V4 or the most

superficial layers of V2 possibly corresponding to the most "elaborate" types of cells in V2. Indeed V2 studies have reported a wide array of cell types with different degrees of complexity, from the simplest neurons being selective to V1-like oriented stimuli [Burkhalter and Essen, 1986; Gegenfurtner et al., 1996], to the most complex ones being selective to V4-like stimuli, *i.e.*, intersections, arcs, circles, texture patterns such as sinusoidal and non-Cartesian gratings [Kobatake and Tanaka, 1994; Hegdé and van Essen, 2000, 2003] and angle stimuli [Ito and Komatsu, 2004]. Recently [Boynton and Hegdé, 2004] suggested that V2 selectivity could be explained by the non-linear combination of V1-like subunits (*i.e.*, precisely what S_2 units do).

Beyond S_2 and C_2 units the same process is iterated once more to increase the complexity of the preferred stimulus at the S_3 level (possibly related to Tanaka's feature columns in TEO, see below), where the responses of a few C_2 units (≈ 100) with different selectivities are combined with a TUNING operation to yield even more complex selectivities. In the next stage (possibly overlapping between TEO and TE), the complex C_3 units, obtained by pooling S_3 units with the same selectivity at neighboring positions and scales, are also selective to moderately complex features as the S_3 units but with a larger range of invariance.

The S_3 and C_3 layers provide a representation based on broadly tuned shape-components. The pooling parameters of the C_3 units (see Appendix A) were adjusted so that, at the next stage, units in the S_4 layer exhibit tuning and invariance properties similar to those of the so-called view-tuned cells of AIT [Logothetis et al., 1995] (see Chapter 3). The receptive field sizes of the S_3 units are about $1.2^o - 3.2^o$ while the receptive field sizes of the C_3 and S_4 units are at least 4^o, *i.e.*, covers the whole stimulus.

Bypass routes

Besides the main route that follows stages along the hierarchy of the ventral stream step-by-step, there exist several routes which *bypass* some of the stages, *e.g.*, direct projections from V2 to TEO [Boussaoud et al., 1990; Nakamura et al., 1993; Gattass et al., 1997] (bypassing V4) and from V4 to TE (bypassing TEO) [Desimone et al., 1980; Saleem et al., 1992; Nakamura et al., 1993]. While the main route constitutes the major source of inputs to IT [Tanaka, 1996], bypass routes remain a significant source of inputs. For instance, TE remains visually connected even after lesions of V4 and/or TE [Buffalo et al., 2005]. In-

deed deficits in discrimination tasks are only moderate after V4 [Schiller and Lee, 1991; Schiller, 1993, 1995; Buffalo et al., 2005] lesions and/or TEO lesions [Buffalo et al., 2005][6]. Also lesion studies have shown that only a limited impairment in fine discrimination tasks [Merigan et al., 1993] was observed after V2 lesions suggesting that routes bypassing V2 (*e.g.*, directly from V1 to V4 [Nakamura et al., 1993]) may play an important role.

In the model, such *bypass* route corresponds to the projections from the C_1 layer to the S_{2b} and then C_{2b} layers (where the 'b' stands for 'bypass'). S_{2b} units combine the response of several retinotopically organized V1-like complex C_1 units at different orientations just like S_2 units. Yet the receptive field size of the corresponding $S2b$ units is larger (2 to 3 times larger) than the receptive field size of the S_2 units. Importantly, the number of afferents to the S_{2b} units is also larger (100 afferents *vs.* 10 only for S_2 units), which results in units which are more selective and more "elaborate" than the S_2 units, yet, less tolerant to deformations. The effect of skipping a stage from C_1 to S_{2b} not only results in units at the C_{2b} level that are more selective than other units at a similar level along the hierarchy (C_3 units), but that also exhibit a lesser range of invariance to positions and scales.

There could be many advantages for a visual system to not only rely on a main (step-by-step) route but also includes bypass routes in parallel. Beyond the obvious robustness to lesions, we speculate that bypass routes may help provide a richer vocabulary of shape-tuned units with different levels of complexity and invariance. Experimentally, we found that, while the level of performance on various categorizations tasks of individual model layers (*i.e.*, C_2 *vs.* C_{2b} *vs.* C_3 units) are fairly similar, their combination, however, provide a significant gain in performance.

A Loose hierarchy

It should be emphasized that the various layers in the architecture – from V1 to TEO – create a large and redundant dictionary of features with different degrees of selectivity and invariance.[7] Yet, it may be advantageous for circuits in later stages (say task-specific circuits in PFC) to have access not only to the highly invariant and selective units of AIT but also to less invariant and simpler units of the V2 and V4 type. Very fine orientation discrimination tasks, for instance, certainly require information from lower levels of the hierarchy such as V1. There might also be high level recognition tasks that benefit from less invariant representations.

For instance, recent work by Wolf & Bileschi has shown that the recognition performance of the model on real-world image databases (see Chapter 4) including different object categories with large variations in shape but limited ranges of positions and scales could be further improved by 1) restricting the range of invariances of the top units and 2) passing some of the C_1 unit responses to the classifier along with the top unit responses [Wolf et al., 2006; Bileschi and Wolf, 2006]. We also found in the animal *vs.* non-animal categorization task in Chapter 5 that the performance is improved with S_4 units that not only receive their inputs from the top C_3 and C_{2b} units but also from low-level C_1 units (with much more limited invariance to position and scale). Finally preliminary computational experiments by Meyers & Wolf suggest for instance that "fine" recognition tasks (such as face identification) may benefit from using units in lower stages (such as C_1 and S_2 units).

Though the present implementation follows the hierarchy of Fig. 2-1, the hierarchy may not be as strict. For instance there may be units with relatively complex receptive fields already in V1 [Mahon and DeValois, 2001; Victor et al., 2006]. A mixture of cells with various levels of selectivity has also commonly been reported in V2, V4 and TEO [Tanaka, 1996]. Hubel & Wiesel already pointed out that the hierarchy of V1 cells they described (*i.e., circular symmetric* $<$ *simple* $<$ *complex* $<$ *hypercomplex*) may not be as strict [Hubel and Wiesel, 1977]. For instance, they (and others) have found cells of the hypercomplex type (showing suppression to elongated bars) but relatively sensitive to the position of the stimulus within their receptive field (like simple cells). In addition, it is likely that the same stimulus-driven learning mechanisms implemented for S_2 units and above (see Section B) operate also at the level of the S_1 units. This may generate S_1 units with TUNING not only for oriented bars but also for more complex patterns (*e.g.*, corners), corresponding to the combination of LGN-like, center-surround subunits.

In cortex there exist at least two ways by which the response from lower stages could be incorporated in the classification process: 1) Through bypass routes (see above, for instance through direct projections between intermediate areas and PFC) and/or 2) by replicating some of the unit types from one layer to the next. This would suggest the existence of cells such as V1 complex cells along with the bulk of more "elaborate" cells in the various stages of visual cortex. We are of course aware of the potential implications of observation (1) for how back-projections could gate and control inputs from lower areas to PFC in order to optimize performance in a specific task (see Chapter 6). From the same point of view,

direct connections from lower visual areas to PFC make sense computationally.

The number of subunits should increase from V1 to IT thus increasing the complexity of the preferred stimulus (this could also produce simple units with broader ranges of invariances for instance). Importantly the size of the receptive fields and the potential complexity of the optimal TUNING grow in parallel. Finally, in the present model implementation, the two layers of simple and complex units alternate from S_1 to S_4 though levels could be conceivably skipped, see [Riesenhuber and Poggio, 1999a]. In particular, consistent with the current model implementation, after sufficient position and scale invariance is obtained (at the C_{2b} and C_3 layers), it is likely that only cells of the S type follow each other.

A.3 Invariant Recognition in IT

IT is believed to play be key in the ability of primates to perform invariant object recognition [Tanaka, 1996]. Based on lesions studies, IT is generally subdivided into two subregions, *i.e.*, posterior (PIT) and anterior (AIT) cortices that are roughly coextensive with, but not identical to, the cytoarchitectonic TE and TEO subdivisions [Logothetis and Sheinberg, 1996]. It has been reported that the selectivity of neurons in TEO and V4 are similar, only the topography is coarser and the receptive fields are larger in TEO [Boussaoud et al., 1991]. Indeed lesions to either V4 and TEO leads to similar deficits [Buffalo et al., 2005] in filtering out distractors at the level of TE. It is therefore likely that there is a lot of overlap between these regions and correspondence with model layers may only be vague. In Chapter 3 we show that the selectivity of the C_2 units seem compatible with the tuning of neurons in V4. The range of invariances of the C_2 units was determined in [Cadieu, 2005] so that it is compatible with the range of invariance of V4 neurons [Pasupathy and Connor, 2001]. It is possible that C_2 units extend in TEO. It might also be that experimentalists were recording from neurons more similar to model S_{2b} and C_{2b} units. Like S_2 units, S_{2b} also receive afferents from V1-like units, yet, because of their position along the hierarchy (pyramidal cells may cover larger extents in space and are more spinous [Elston, 2003]), they receive a larger number of retinotopically organized afferents, have larger receptive fields and are therefore more selective (in the model S_{2b} units receive inputs from 100 afferents whereas S_2 units receive only 10 afferents).

As one progresses from the posterior part of TEO to the most anterior part of TE, the topography is almost completely lost and the size of the receptive field sizes increase significantly (receptive field sizes can be as small as $1.5^o - 2.5^o$ in TEO [Logothetis and Sheinberg, 1996]) so as to cover large parts of the visual field (up to $30^o - 50^o$ [Boussaoud et al., 1991; Tanaka, 1993] in TE). This increase in the receptive field sizes is also accompanied by a significant increase in the complexity of the preferred stimulus (from simple stimuli to complex shapes such as faces or hands), [see Logothetis and Sheinberg, 1996]. It is therefore likely that S_3 units may overlap between TEO and TE (S_3 units receive inputs from several V4-like cells with limited invariance). C_3 units that cover a range of invariance of about $\pm 2^o$ in translation and ≈ 2 octaves in scale are likely to be found more anteriorly in TE (see Chapter 3). Neurons that respond to parts of objects, *e.g.*, the eyes [Perrett et al., 1982, 1992], may correspond to S_{2b} or C_{2b} units (depending on their invariance properties), while neurons that require the simultaneous presentation of multiple parts of a face [Perrett and Oram, 1993; Wachsmuth et al., 1994] may correspond to S_3, C_3 or maybe even S_4 units (see Section C).

As we discussed in Chapter 1, one unit in the model may be described best by a *computational module*, composed of a set of n *equivalent* neurons, *i.e.*, receiving the same inputs. It is therefore not surprising to find columns of features in IT [Fujita et al., 1992; Wang et al., 1996; Tanaka, 1996, 1997; Wang et al., 1998; Tanaka, 2003]. A natural question to ask is to compare the size of the vocabulary of the shape tuned units used in the model at the level of IT with the number of columns of features found in IT. In [Serre et al., 2005b] (see also B), we evaluated the categorization performance of a linear classifier (SVM and Ada-Boost) that uses as an input, a subset of the dictionary of shape component units from the model. While the performance increases monotonically when increasing the number of features used, we found that the level of performance reaches a ceiling between $1,000 - 5,000$ features. The comparison with cortex becomes quiet remarkable: Tanaka and colleagues [Fujita et al., 1992] estimated the number of columns in TEd to be $\approx 2,000$!

B Learning a Dictionary of Shape-Components from Natural Images

B.1 On Learning Correlations

Various lines of evidence suggest that visual experience – during and after development – together with genetic factors determine the connectivity and functional properties of cells in cortex (see Chapter 1). In this work, we assume that learning plays a key role in determining the wiring and the synaptic weights for the model units.

More specifically, we suggest that the TUNING properties of simple cells – at various levels in the hierarchy – correspond to learning which combinations of features appear most frequently in images. This is roughly equivalent to learning a dictionary of image patterns that appear with higher probability. The wiring of the S layers depends on learning correlations of features in the image **at the same time** (*i.e.,* for S_1 units, the bar-like arrangements of LGN inputs, for S_2 units, more complex arrangements of bar-like subunits, *etc*).

The wiring of complex cells, on the other hand, may reflect learning from visual experience to associate frequent transformations in time – such as translation and scale – of specific complex features coded by simple cells. The wiring of the C layers reflects learning correlations **across time**, *e.g.,* at the C_1 level, learning that afferent S_1 units with the same orientation and neighboring locations should be wired together because, often, such a pattern changes smoothly in time (under translation) [Földiák, 1991].

Thus learning at the S and C levels is learning correlations present in the visual world. At present it is still unclear whether these two types of learning require different types of synaptic "rules" or not.

B.2 The Learning Rule

The goal of this learning stage is to determine the selectivity of the S units, *i.e.*, set the weight vector w (see Eq. 1.2) of the units in layers S_2 and higher. More precisely, the goal is to define the basic types of units in each of the macro-columns (see Fig. 2-3). We suggest that these macro-columns (or feature-maps) constitute a basic dictionary of shape-components with units that are tuned to *image-features* that occur with high probability in

natural images.

This is a very simple and natural assumption. Indeed it follows a long tradition of researchers that have suggested that the visual system, through visual experience and evolution, may be adapted to the statistics of its natural environment [Attneave, 1954; Barlow, 1961; Atick, 1992; Ruderman, 1994] (see also [Simoncelli and Olshausen, 2001] for a recent review). For instance, [Attneave, 1954] proposed that the goal of the visual system is to build an efficient representation of the visual world and [Barlow, 1961] emphasized that neurons in cortex try to reduce the redundancy present in the natural environment.

More recently, theoretical studies have shown that receptive fields that resemble cells in primary visual cortex can be learned (through non-biological optimization techniques) based on several learning principles, *e.g.*, sparseness [Olshausen and Field, 1996] (minimizing the number of units active for any input), statistical independence [Hyvärinen and Hoyer, 2001] or even temporal continuity and slowness [Wiskott and Sejnowski, 2002; Körding et al., 2004; Berkes and Wiskott, 2005]. Regularities in natural visual scenes may also provide critical cues to the visual system to solve specific tasks [Richards et al., 1992; Knill and Richards, 1996; Callaway, 1998b; Coppola et al., 1998] or even provide a teaching signal [Barlow, 1961; Sutton and Barto, 1981; Földiák, 1991] for learning with no supervision.

In the model, we assume that this learning stage is unsupervised and may occur during a developmental-like learning stage . It is likely that new features may be learned after this initial learning stage during adulthood (certainly at the level of IT [Logothetis et al., 1995; Kobatake et al., 1998; Booth and Rolls, 1998; Sigala and Logothetis, 2002; Baker et al., 2002] and even in intermediate [Yang and Maunsell, 2004; Rainer et al., 2004] and lower areas [Singer et al., 1982; Karni and Sagi, 1991; Yao and Dan, 2001; Schuett et al., 2001; Crist et al., 2001], see Chapter 1), possibly in a category- or task-specific manner as people become experts for specific recognition problems[9] Yet, our results suggest that it is possible to perform robust invariant object recognition from a generic set of shape-tuned units learned with no supervision [Riesenhuber and Poggio, 1999a] from a general set of natural images unrelated to any categorization task.

Learning in the model is sequential, *i.e.*, layers are trained one after another (all images from the database are presented during the training of each individual layers) starting from the bottom with layers S_2 and S_{2b} and then progressing to the top with layer S_3.[10]

Figure 2-7: Sample natural images used to expose the model and learn the generic dictionary of shape-components from V2 to IT.

During this developmental stage, the weights $(\mathbf{w^1}, \ldots, \mathbf{w^n})$, *i.e.*, the preferred stimulus, of the S units within each mini-column which are shared across all macro-columns in the layer, are learned sequentially starting with $\mathbf{w^1}$ and ending with $\mathbf{w^n}$. At the k^{th} image presentation, one macro-column (which corresponds to a particular portion of the visual field and scale) is selected (at random) and unit $\mathbf{w^k}$ from this macro-column is *imprinted*, *i.e.*, the unit stores in its synaptic weights the current pattern of activity from its afferent inputs in response to the part of the natural image that fell within its receptive field. This is done by setting $\mathbf{w^k}$ to be equal to the current pattern of pre-synaptic activity $\mathbf{x}$.[11] As a result, the image patch $\mathbf{x}$ that falls within the receptive field of unit $\mathbf{w^k}$ becomes its preferred stimulus. After this imprinting process, the unit is mature.

During this learning stage, we also assume that the image moves (shifting and looming) so that the selectivity of the unit $\mathbf{w^k}$ is generalized within the same mini-column to units at different scales (looming) and across macro-columns (shifting) to units at different locations in the visual field with a generalized Hebbian rule [Földiák, 1991] (see Subsection B.1). Note that we did not implement this generalized Hebbian rule, and simply "tiled" unit $\mathbf{w^k}$ across scales and positions.

Learning all n S unit types within one layer thus requires exposing the model with n images. The database of images we used contained a large variety of natural images

collected from the web (including landscapes, street scenes, animals, *etc*), see Fig. 2-7 for examples. The dictionary of shape components learned during this developmental learning stage is generic in that, as we show in Chapter 4, the same basic dictionary can be used for the robust and invariant recognition of many different object categories. Afterward, only the task-specific circuits from IT to PFC required learning for the recognition of specific objects and object categories.

C Building Task-Specific Circuits from IT to PFC

We assume that a particular *program* or *routine* is set up somewhere beyond IT (possibly in PFC [Scalaidhe et al., 1999; Freedman et al., 2002, 2003; Hung et al., 2005] but the exact locus may depend on the task). In a passive state (no specific visual task is set) there may be a default routine running (perhaps the routine: *what is there?*). Here we think of this routine as a particular PFC-like *classification* unit which combines the activity of a few hundred S_4 units tuned to examples of the target object (as well as distractors). While learning in the model from S_2 to S_4 is stimulus-driven, the PFC-like *classification* units are trained in a *supervised way*.

S_4 view-tuned units: Consistent with a large body of data that suggests that the selectivity of neurons in IT depends on visual experience (particularly training) and that the corresponding learning-related changes may be very fast (see Chapter 1), we assume that, when a new task is being learned, S_4 units which correspond to the so-called view-tuned cells of AIT, become selective to specific examples of the training set (*e.g.*, views of the target objects). This is in good agreement with the specificity of IT neurons to certain object views or lighting conditions. For example, [Logothetis et al., 1995] found that after training monkeys to perform an object recognition task with isolated views of novel three-dimensional paperclip objects, the great majority of neurons selectively tuned to the training objects were view-tuned to one of the training objects. About one tenth of the tuned neurons were view-invariant, consistent with an earlier computational hypothesis [Poggio, 1990].

In the present model implementation, during the training of the task-specific circuits, a small fraction ($\approx 25\%$) of the training set of objects (for instance examples of cats and dogs for a cat *vs.* dog discrimination task) is stored at the level of the S_4 units. Just like units in

lower stages become tuned to patches of natural images, S_4 units become tuned to views of the target object by storing in their synaptic weights the precise pattern of activity from their afferents during a presentation of a particular exemplar.

It is important to point out that, consistent with the notion of a loose hierarchy described in Section A, we found that, while using the S_4 stage improves the overall performance of the model, reasonably good results can be obtained without this S_4 stage. Robust invariant recognition performance can be obtained by a linear classifier that uses part of the dictionary of shape-components directly. In particular, in [Serre et al., 2005b], we showed that the same linear SVM classifier as in [Hung et al., 2005] (see above) actually competes with some of the best computer vision systems when using the model C_{2b} features as inputs (see also Chapter 4).

PFC-like classification units: The proposal that classification tasks could be performed in cortex by a linear classifier in higher areas (such as PFC) that integrates the activity of a few hundred neurons from the example-based view-tuned units in the S_4 layer (corresponding to cells in IT) was originally formulated by [Poggio and Edelman, 1990] to explain view-invariant recognition and later extended in [Riesenhuber and Poggio, 2000]. The concept of a linear classifier that takes its inputs from a few broadly tuned example-based units is a powerful learning scheme that is closely related to Radial Basis Function (RBF) networks, which is among the most powerful in terms of learning to generalize [Poggio and Girosi, 1990; Poggio and Smale, 2003]. Computer simulations have shown the plausibility of this scheme for visual recognition and its quantitative consistency with many data from physiology and psychophysics [Poggio and Bizzi, 2004]. In particular, [Poggio and Bizzi, 2004] suggested that the broad tuning of units in IT may be key in providing good generalization properties to classification units beyond IT (for instance to limited changes in pose, illumination, etc).

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

In the model, the response of a PFC-like *classification* unit with input weights $\mathbf{c} = (c_1, \ldots, c_n)$ is given by:

$$f(\mathbf{x}) = \sum_i c_i K(\mathbf{x^i}, \mathbf{x}) \quad \text{where} \quad K(\mathbf{x^i}, \mathbf{x}) = \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j^i - x_j)^2\right) \tag{2.1}$$

characterizes the activity of the i^{th} S_4 unit, tuned to the training example $\mathbf{x^i}$, in response to the input image $\mathbf{x}$ and was obtained by replacing the weight vector $\mathbf{w}$ in Eq. 1.2 by the training example $\mathbf{x^i}$, *i.e.*, $\mathbf{w} = \mathbf{x^i}$. The superscript i indicates the index of the image in the training set and the subscript j indicates the index of the pre-synaptic unit. Supervised learning at this stage involves adjusting the synaptic weights $\mathbf{c}$ so as to minimize the overall classification error on the training set E, such that:

$$E = \sum_{i=1}^{l} ||f(\mathbf{x^i}) - y^i||^2 + R(f). \tag{2.2}$$

where y^i corresponds to the true label $(0-1)$ of the training example $\mathbf{x^i}$ and $f(\mathbf{x^i})$ corresponds to the response (or prediction) of the classification unit to example $\mathbf{x^i}$.
$R(f) = \lambda||f||$ is a regularization term which enforces a smoothness criterion on the function f and which could be omitted for simplicity. Neglecting $R(f)$ for simplicity, minimizing E corresponds to minimizing the classification error on the training set.[12]

In the current model implementation, one PFC-like classification unit is trained for each categorization task. For instance, for the model to be able to recognize all objects from the *CalTech-101* database (see Chapter 4), 101 different PFC-like units f^h with different synaptic weights $\mathbf{c^h}$ are being trained, one for each of 101 objects *vs.* the rest. For a new image presentation, the label h of the PFC-like unit f^h with the maximal output across all PFC-like units is considered the final model response. Alternatively the model can be tested in an object present/absent task like the animal present/absent task described in Chapter 5 by comparing the output of one PFC-like unit $f^o \in [0, 1]$ trained on an animal *vs.* non-animal to a fixed threshold $f^o \lessgtr \theta$.

One may raise the concern that training the model for all possible categories in the world would lead to an intractable number of units in the model. Yet, this is not the case. First, as we discussed earlier, it is possible to skip the S_4 stage and maintain a high level of performance (the S_4 stage may only correspond to objects for which we are **expert or highly familiar** with, *e.g.*, faces or body parts, cars, places, etc. Training the model to recognize a plausible number of discriminable objects (i.e., probably no more than $30,000$ [Biederman, 1987]), would add $\sim$ 3 million S_4 units (assuming a realistic $\sim$ 100 S_4 per

class). The number of neurons in AIT is ~ 15 million in each hemisphere [J. DiCarlo, Pers. Comm.]. At the PFC-like level the number of *classification* units required would be very small $\sim 100,000$. It is also very likely that in cortex the same units would be involved in different categorization tasks [E. Miller, Pers. Comm.]. Key in the model is the use of a generic dictionary of shape components common to most object and therefore prevent an explosion in the number of units needed for recognition.

In summary: To perform a new categorization task (for instance a cat *vs.* dog categorization task [Freedman et al., 2001]), the model is trained in a supervised way by:

1. Storing part ($\approx 25\%$) of the training examples (*i.e.*, images of cats and dogs, the exact number may vary) at the level of the S_4 units;

2. Training the task-specific circuits from IT to PFC by setting the synaptic weights $\mathbf{c}$ of an PFC-like *cat* vs. *dog classification* denoted $f^o \in [0,1]$ so as to minimize the classification error on the training set;

3. To evaluate the model performance, a model prediction is obtained for each image $\mathbf{x}$ from a test set of images (disjoint from the training set) by thresholding the unit response $f^o(\mathbf{x}) \lessgtr \theta$. An average classification error is computed by counting the number of mismatches between the model prediction and the true label of all images.

On the difference between S_4 and categorization units: Would a physiologist be able to tell apart a S_4-like cell from a classification cell in cortex? First, based on its dense connectivity with other cortical and subcortical areas as well as its potential involvement in representing stimulus-reinforcement associations and reward [Rolls, 2000; Tremblay and Wolfram, 2000], PFC is likely to be a prime location for the classification cells. Conversely, we expect most S_4-like (view-tuned) cells to be found in AIT/TE and STS (see Chapter 1). An objective criterion may require a measure of category selectivity, *e.g.*, the *category index* used by [Freedman et al., 2002], which has been shown to be higher for neurons in PFC than in IT in monkeys that have been trained to discriminate between cats and dogs. As a result, we predict that face cells found in the inferior prefrontal cortex by [Scalaidhe et al., 1999] may correspond to face classification cells possibly involved in different tasks (*e.g.*, face categorization, face identification, expression recognition, gender recognition, *etc*).

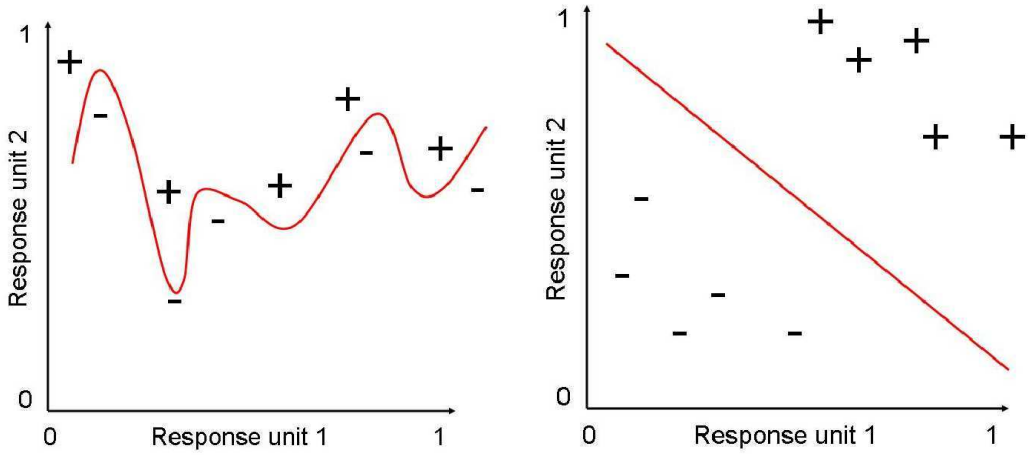
D Discussion

D.1 On Learning Good Features

Building better image representations has been a major effort among computer vision scientists over the past decade [Leung et al., 1995; Mohan et al., 2001; Ullman et al., 2002; Heisele et al., 2002; Lowe, 2004] and the emerging picture is that better recognition systems will require better features rather than better (more complex) classification algorithms.

This principle is illustrated in Fig. D.1. Imagine you are trying to build a classifier to read out object categories from the output of two units (corresponding to neurons in IT for example and denoted *unit 1* and *unit 2*). The responses of these two units to various examples of the target object (*e.g.*, under different views, illuminations, *etc*) are characterized by $(+)$ and to distractors (*e.g.*, examples from other object categories as well as background images) by $(-)$. In both panels 2-8(a) and 2-8(b), it is possible to find a separation (the red line indicates one such possible separation) between the two sets of data-points. Yet, statistical learning theories [Vapnik, 1995] teaches us that the representations provided by the units in the two panels are not equal: The representation provided by the two units in panel 2-8(b) is far superior to the one provided by the units in panel 2-8(a). The reason is that the classifier in panel 2-8(b) is much simpler than the classifier in panel 2-8(a) (a rough estimate of the *complexity* of a classifier is given by the number of wiggles of the separation line). Learning the separation in panel 2-8(b) tends to be faster and require much less training examples. Additionally because the data-points in panel 2-8(b) lie further away from the separation, it is guaranteed to generalize better to new previously unseen examples than in panel 2-8(a).

At the neural level, the difference between the two panels could result from the tuning of the two units. In panel 2-8(b), both units tend to be more selective for the target object than the set of distractors. Fig. 2-9 illustrates this point from empirical simulations with two versions of the model evaluated on a difficult face detection task: one model implementation corresponds to the original model with "hardwired features" [Riesenhuber and Poggio, 1999a] (simple 2×2 combinations of orientations), the other corresponds to a model implementation that uses "learned" features that correspond to prototypical parts of the target object class (learned with k-means from the positive training set of images, see [Serre et al., 2002] for details). This "learned" feature version of the model correspond to



(a) Bad representation (b) Good representation

Figure 2-8: Comparing the quality of the representation provided by two units, *e.g.*, the normalized response of these two units to the presentation of examples of target objects (+) and distractors (−). From the statistical learning theory perspective [Vapnik, 1995], the representation provided by the units in **b)** is far superior to the one in **a)**. At the unit level, the difference between **a)** and **b)** is that units in **b)** are more selective to the target object than in **a)**. Such increased selectivity may result from learning a better feature representation and may provide faster learning and better generalization to new examples.

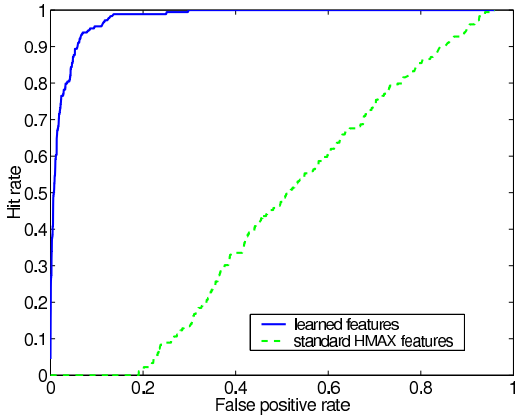


Figure 2-9: Face detection in natural images: Comparison between "learned features" [Serre et al., 2002] and "standard HMAX features" [Riesenhuber and Poggio, 1999a]. The gain in performance after introducing learning in the model is significant. Also note that the "learned feature" version corresponds to an earlier implementation of the model [Serre et al., 2002] which under-performs significantly the model implementation described in this thesis.

an earlier implementation of the current model described here. The gain in performance with the addition of the learning stage is drastic. While the "hardwired" features were able to support the invariant recognition of simple object such as paperclips or even synthetic face examples in the absence of clutter (see [Serre et al., 2002], they failed to handle a face categorization task in natural images [Serre et al., 2002] (*i.e.*, with clutter and large variations in shape and illumination).

D.2 One Basic Dictionary of Features for Different Recognition Tasks

As we later confirm in Chapter 4, the same circuits and mechanisms that we have described in this Chapter, can support the robust and invariant recognition of many different object categories (including faces as well as other objects). Additionally very recent results by Meiers & Wolf (unpublished) also suggest that the model can perform face identification (*i.e.,* at the subordinate level) very well. Altogether this suggests that view-tuned cells in AIT could support the recognition of a wide range of object categories at different levels of categorization. This may suggest, as discussed by [Riesenhuber and Poggio, 2000] and contrary to other proposals [Kanwisher, 2003], that there is no need for special processing or computational mechanisms to support the recognition of different classes of objects or different levels of categorization. [13]

Expert Features: Interestingly the same basic *generic* dictionary of shape components learned from a set of natural images, unrelated to any categorization tasks, is able to handle the robust and invariant recognition of multiple object categories. This is in agreement with the observation that following familiarization to a new object category, rapid changes may occur in higher brain areas (presumably at the level of the task-specific circuits, see Chapter 1). Yet, there is some evidence suggesting learning-related changes in lower areas (see Chapter 1 for a review) suggesting that the dictionary of shape-components may not be *static* and could undergo changes.

Indeed, with simulations, we have found that the performance of the model could be further increased with a more specific dictionary of shape-components. For instance, Fig. 2-10 shows the performance of a linear classifier that uses two different dictionaries of features: a *generic* (called "universal", X-axis) dictionary of features learned from a set of natural images unrelated to any categorization task (see Section B) and an *object-specific* (denoted "specific", Y-axis) dictionary of features learned from a set of images that belong to the target set. Each data-point in the figure represents the performance of the *generic vs.* the *expert* set of features for each one of the 101 different object categories available for testing (see Chapter 4) and for a specific number of examples available for training.

Interestingly with very few training examples (*i.e.*, , less than 6 examples, blue, green and red dots), the performance of the *generic* feature set is higher (probably due to over-fitting of the *object-specific* set that learns both the features and the discrimination function

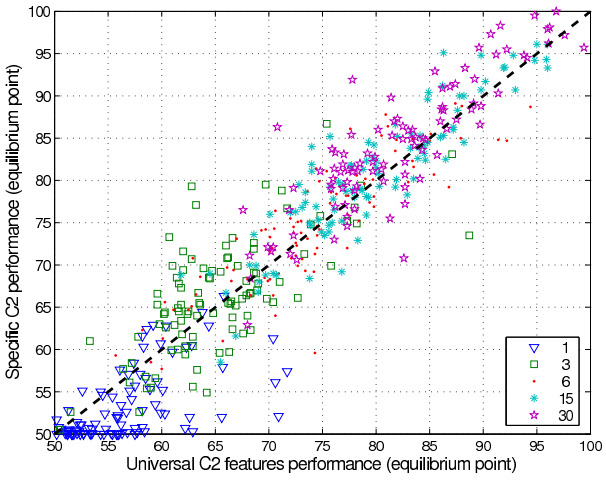


Figure 2-10: *Expert vs. generic* features on the *CalTech-101* object database (see Chapter 4). For each of the 101 object category available for testing and for different number of examples available for training, we compare the performance of a linear classifier that uses a *generic* ("universal", X-axis) *vs.* an *expert* ("specific", Y-axis) set of features. With very few training examples, the performance of the *generic* feature set is higher. Yet, as the number of examples available for training increases the performance of the *expert* set slowly takes over. Reproduced from [Serre et al., 2006b].

from the same training set of images). Yet, as the number of examples available for training increases (baby-blue and violet), the *expert* set starts to slowly outperform the *generic* set. The figure is reproduced from [Serre et al., 2006b] (see [Serre et al., 2006b] for details on the experimental procedure). The *generic* set of features may correspond to learning during development (see Section B) while the *expert* set may correspond to an expert learning stage that occurs later during adulthood. For instance, it has been reported that neurons in AIT becomes tuned to parts of the target objects after extensive training [Logothetis et al., 1995; Sigala and Logothetis, 2002; Baker et al., 2002]. The building of a more specific dictionary of shape-components may be related to the acquisition of expertise described in psychology [Schyns et al., 1998; Williams et al., 1998].

We have experimented with a simple biologically plausible algorithm for learning such *expert* set of features [Serre and Poggio, 2004]. The algorithm uses the temporal association between successive image frames that contain examples of the target object undergoing a transformation. After the presentations of several frames, the learning rule produces a stable representation that is invariant to the transformations undergone by the target object (*e.g.*, clutter, illumination, intra-class variations, *etc*). The proposed algorithm extends previous work [Földiák, 1991; Perrett et al., 1984; Hietanen et al., 1992; Wallis et al., 1993; Wachsmuth et al., 1994; Wallis and Rolls, 1997; Elliffe et al., 2002; Einhäuser et al., 2002; Wiskott and Sejnowski, 2002; Spratling, 2005] that have used the same principle of temporal continuity to learn invariances to position, pose or illumination. We have successfully

used the learning rule to learn critical features for the recognition of biological motion [Sigala et al., 2005] in a model of the dorsal pathway [Giese and Poggio, 2003]. We have also applied it to learn a set of features to be used by top-down attentional circuits [Walther et al., 2005].

Such learning rule finds partial support from psychophysics [Wallis and Bülthoff, 2001] and seems consistent – as pointed out by Stryker [Stryker, 1991; Földiák, 1998; Giese and Poggio, 2003] – with a study by Myashita, who showed, that training a monkey with a fixed sequence of image patterns lead to a correlated activity between those same patterns during the delayed activity [Miyashita, 1988].

D.3 What is the Other 99% of Visual Cortex Doing?

As described in Table 2.1[14], the model contains on the order of 10 million units (these bounds are computed using reasonable estimates for the S_4 receptive field sizes and the number of different types of simple units in all S layers). This number may need to be increased by no more than one or two orders of magnitude to obtain an estimate in terms of biological neurons – based on the circuits described in [Serre et al., 2005a]. This estimate results in about $10^8 - 10^9$ actual neurons, which corresponds to about 0.01% to 1% of visual cortex (based on 10^{11} neurons in cortex [Kandel et al., 2000]). This number is far smaller than the proportion of cortex taken by visual areas ($\sim 50 - 60\%$).

We shall emphasize that, even though this number was computed for a version of the model trained to perform a single (binary) animal *vs.* non-animal classification task – because the same basic dictionary of shape-tuned units (*i.e.*, from S_1 up to S_4) is being used for different recognition tasks – this number would not differ significantly for a more realistic number of categories. In particular, training the model to recognize a plausible number of discriminable objects (*i.e.*, probably no more than $30,000$ [Biederman, 1987]), would add only an extra 10^7 S_4 units.

Obviously our model does not constitute a complete model of visual cortex. In particular the calculation from Table 2.1 only takes into accounts units from the ventral stream. Also several processing channels such as color, motion and stereo would have to be incorporated. So far the biophysical simulations to implement the key operations have been performed with very few inputs [Serre et al., 2005a] and it is possible that the key computations may have to be broken down into several sub-computations to handle larger number

Layers	Number of units
S_1	1.6×10^6
C_1	2.0×10^4
S_2	1.0×10^7
C_2	2.8×10^5
S_3	7.4×10^4
C_3	1.0×10^4
S_4	1.5×10^2
S_{2b}	1.0×10^7
C_{2b}	2.0×10^3
Total	2.3×10^7

Table 2.1: Number of units in the model. The number of units in each layer was calculated based on the animal *vs.* non-animal categorization task presented in Chapter 5, *i.e.*, S_4 (IT) receptive fields (RF) spanning only 4.4^o of visual angle (160×160 pixels, probably not quite matching the number of photoreceptors in the macaque monkey in that foveal area of the retina) and about $2,000$ types of units in each S_2, S_{2b} and S_3 layers.

of afferents. Yet taken all of these limitations into account and assuming that we need to increase our estimate of the number of neurons by one order of magnitude, it remains that the model can categorize visual object with no more than 10% of visual cortex.

Note that with large scale neural architectures such as *blue brain*, it is expected that we will soon be able to simulate $\approx 10^8$ single-compartment neurons[15] and therefore simulate the model with more detailed units. In particular, to make the simulations computationally tractable, the model presented here only uses a *static* approximation of the two key computations, *i.e.*, TUNING and MAX operation. As discussed in Chapter 1, a better description of the two key computations will involve biophysical micro-circuits (see Fig. 1-8).

Notes

[1]By convention, 1^o of visual angle in the model corresponds to 36×36 pixels in the input image (see Appendix A).

[2]Our model of simple cells as Gabor filters, applied directly on the raw pixel image, is an obvious oversimplification. Yet, as we show in Chapter 3, the corresponding S_1 units constitute a good *phenomenological model* of simple cells and account well for the the tuning properties of cortical simple cells. A more realistic implementation would correspond to the combination of center-surround (ON and OFF) ganglion cell receptive fields with

the TUNING operation described in Eq. 1.3. Feedforward push-pull mechanisms [Ferster and Miller, 2000; Miller, 2003; Hirsch, 2003] which combine a balance of feedforward ON excitation with feedforward OFF inhibition (or vice-versa) could be implemented by the numerator of Eq. 1.3. In principle, the denominator of Eq. 1.3 could provide contrast adaptation through feedforward shunting inhibition (from inhibitory interneurons in layer IV) or even sharpening of the orientation through recurrent shunting inhibition [Sompolinsky and Shapley, 1997] (from cortical cells at other preferred orientations – which have zero weights in the numerator and therefore do not participate in shaping the classical receptive field of the S_1 unit).

[3]When parameterizing S_1 units we tried to account for an observation about cortical V1 cells which is that larger cells tend to be tuned to lower spatial frequencies and vice-versa, see Chapter 3 and Appendix A.

[4]As we discussed in Chapter 1, units in the model are more likely to correspond to *computational modules* in cortex, *i.e.*, ensemble of n *equivalent* cells with the same inputs rather than single neurons. Each mini-column in the model is thus composed of several modules at different scales. We suggested earlier that the number n of cells in each computational module may decrease along the hierarchy. Additionally we suggest that both the number of scales in each mini-column as well as the number of macro-columns may also decrease (with cells becoming more and more invariant to scale and position). Alternatively we propose that the number of mini-columns within each macro-columns may increase (from only 8 types of units at the S_1 level (4 orientations and 2 phases) to about $1,000$ types of units in higher stages).

[5]The sampling is reduced from S_1 to C_1 layers. From 17 different scales at the S_1 level, the scale space is reduced to only 8 scales at the C_1 level (with broader frequency tuning). Similarly a large downsampling is performed over positions (see Appendix A). Yet, in all layers, there is a high degree of overlap between units.

[6]Both V4 and TEO could also be bypassed through structures that have not been yet incorporated in the model, *e.g.*, subcortical areas [Baleydier and Morel, 1992; Webster et al., 1994b], the superior temporal sulcus denoted *STS* in Fig. 1-1 [Saleem et al., 1996], the

perirhinal (*35* and *36*) and parahippocampal areas denoted *TF* and *TH* in Fig. 1-1 [Webster et al., 1991; Horel, 1992], and through the parietal and frontal cortex [Webster et al., 1994a] from V2 [Boussaoud et al., 1990].

[7]Several researchers have emphasized the computational constraints on invariant feature-based representations of the kind used in the model and the difficult trade-off between selectivity and invariance in achieving invariant recognition [von der Malsburg, 1981, 1995, 1999; Riesenhuber and Poggio, 1999b,a; Ullman and Soloviev, 1999; Ullman et al., 2002; Mel and Fiser, 2000; Stringer and Rolls, 2000; Amit and Mascaro, 2003]. Loosely speaking, the simpler the feature-detector, the more likely it is to produce false-alarms. As a result, simple feature-detectors are only useful with a limited range of invariance. To build more invariant representations it is important to rely on more complex feature detectors. Accordingly it has been suggested that the gradual parallel increase in the invariance properties and the preferred stimulus of cells along the ventral stream is a result of this invariance-selectivity trade-off [Riesenhuber and Poggio, 1999b,a; Mel and Fiser, 2000]. In particular, [Mel and Fiser, 2000] used an analytical model (in the domain of text) to study this design trade-off and the susceptibility of such system to false-positive recognition errors.

[8]A classic misconception is that invariant feature-based representations are insensitive to image scrambling. This would only be true for a representation that relies on very few, non-overlapping parts. But as feature-detectors become more numerous and start to overlap, *i.e.*, the representation becomes more redundant, then scrambling the image disrupts at least part of the features (the precise number of features disrupted likely depends on the relative size of the features and the scrambling procedure), see [Riesenhuber and Poggio, 1999a] for quantitative results.

[9]We recently proposed a biologically plausible learning rule for selecting the most informative features about an object class, *i.e.*, features that repeat across different images of the same objects. We have simulated a simplified version of Földiak's trace rule to generate units that become tuned to complex features of images [Serre and Poggio, 2004]. After presentation of many natural images, the units become tuned to complex features – for instance of face-components – if a sequence of face images (in the presence of background) is presented (in general, objects are not at the same position and scale). Learning is task-

independent and simply relies on temporal continuity (*e.g.*, the same object being present during a temporal sequence of images).

[10]A more realistic implementation would require a continuous learning of all the layers with fast time constants in the top layers and increasingly slower time constants from top to bottom.

[11]A more biologically plausible version of this rule would involve mechanisms such as LTP [Markram et al., 1997; Bi and Poo, 1998; Abarbanel et al., 2002; van Rossum et al., 2000].

[12]There are several ways to solve Eq. 2.2. In this thesis, we computed a simple linear least-square fit solution using Matlab© (The MathWorks, Inc) left division operation for matrices. We also obtained very similar results with a stochastic gradient using weight perturbations (only it takes longer to train). Such approach is simpler to implement in neural hardware, as it does not attempt to solve the global optimization problem. Instead, at each iteration, a small step is taken in a random direction in the parameter space (*i.e.*, a small noise vector ξ is added to the weight vector $\mathbf{c}$ to yield the new weight vector $\mathbf{c}^* = \mathbf{c} + \xi$) yielding the new classification function f^*. The error $||f^*(\mathbf{x^k}) - y^k||^2$ on the current training image $\mathbf{x^k}$ is compared to the error performed by the classification unit before the random step was taken $||f(\mathbf{x^k}) - y^k||^2$. If the error decreases with the update, the update is maintained, else a step in the opposite direction is taken. Interestingly, this simple algorithm is guaranteed to minimize the gradient ∇E on average. Such approach is analog to a Darwinian evolution for learning in brains and provides a computational role for the randomness of synaptic transmission in cortex [Seung, 2003].

[13]It would be interesting to perform an fMRI experiment on the model. Based on the observation that cells which become selective to a new object tend to form clusters of cells [Erickson and Desimone, 1999] with similar selectivities (see also Chapter 2 and Tanaka's feature columns in TEO), it is likely that S_4-like cells in AIT, tuned to different examples of the same object, would be neighboring in space. As a result, one may speculate that such clusters of units in the model could potentially appear to form "areas", *e.g.*, a *Face Area* [Kanwisher et al., 1997] when the model is trained with faces or a *Body Area* [Downing

et al., 2001] when trained with body parts, *etc* .

[14]The numbers in Table 2.1 are only "suggestive". In particular, no effort has been made in using realistic numbers in the relative proportion of units in different stages.

[15]`http://bluebrainproject.epfl.ch`

[16]It is interesting to point out that the approach developed in the model bares some connections with other non-biological computer vision systems. For instance, the dictionary of shape-components represented from the S_2 to the S_4 layers and the resulting hierarchy of unit selectivities is somewhat similar to features such as components [Mohan et al., 2001; Heisele et al., 2001a, 2002], parts, [Burl et al., 1998; Weber et al., 2000a; Fergus et al., 2003; Fei-Fei et al., 2004], fragments [Ullman and Soloviev, 1999; Ullman et al., 2002], codewords [Jurie and Triggs, 2005], keypoints [Lowe, 2004] and bags of features [Csurka et al., 2004] in computer vision. Yet, contrary to these computer vision systems that learns new features for each new object class to be learned, the dictionary of features used by the model is *generic*, *i.e.*, it can support several different recognition tasks and in particular recognition of many different object categories.

[17]Our model implementation may also provide a quantitative framework to the very vague notion of pre-attentive vision and "unbound" features in cognitive science [Treisman and Gelade, 1980; Wolfe and Bennett, 1997; Evans and Treisman, 2005].

Chapter 3

Comparison with Neurons

We have described earlier in Chapter 2 the general architecture, the organization as well as the main functional primitives of the model. Here we show that layers of the model can be mapped to cortical areas. In particular, we describe a quantitative comparison between the model and data from V1, V4 and IT (TE). While some of the model parameters were manually adjusted so that the lower stages in the model (*i.e.*, S_1, C_1) as well as the top layer (*i.e.*, S_4) match the tuning properties of cells in V1 and TE respectively (see Section A and C), no parameters were fit in intermediate layers of the model. It is therefore quiet remarkable that the model units in the C_2 stage, as shown in Section B, agrees with V4 data.

A V1 and the Model

The parameters of the S_1 and C_1 units were manually adjusted so that their receptive field sizes, frequency tuning and orientation bandwidth span the range of V1 parafoveal simple and complex cells when probed with standard stimuli (*i.e.*, single bars, edges and gratings). Because the tuning properties of cortical cells vary widely along a continuum (see [Hubel and Wiesel, 1968; Schiller et al., 1976b,c,e,a; DeValois et al., 1982b,a]), it would be very difficult to quantitatively account for the whole population of cells. Instead we tried to generate a population of units that accounts for the *bulk* of cells in primate striate cortex. As illustrated in Table 3.1, the population of S_1 and C_1 units is able to capture the tuning properties of typical simple and complex cells.

A.1 Methods

Orientation Tuning

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

Spatial Frequency Tuning

The spatial frequency selectivity of each model unit was assessed by sweeping sine wave gratings at various spatial frequencies over the receptive field of the unit. For each grating frequency, the maximal unit response was used to fit a tuning curve and the *spatial frequency selectivity bandwidth* was calculated as in [DeValois et al., 1982a] by dividing the frequency score at the high crossover of the curve at half-amplitude by the low crossover at the same level. Taking the $\log_2$ of this ratio gives the bandwidth value (in octaves):

$$\text{bandwidth } = \log_2 \frac{\text{high cut}}{\text{low cut}} \tag{3.1}$$

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

$$\text{selectivity index} = \frac{\text{high cut}}{\text{low cut}} \times 100 \tag{3.2}$$

	Receptive field sizes		
	Model	Cortex	References
simple cells	$0.2^o - 1.1^o$	$\approx 0.1^o - 1.0^o$	[Schiller et al., 1976e; Hubel and Wiesel, 1965]
complex cells	$0.4^o - 1.6^o$	$\approx 0.2^o - 2.0^o$	
	Peak frequencies (cycles /deg)		
	Model	Cortex	References
simple cells	range: $1.6 - 9.8$	bulk $\approx 1.0 - 4.0$	[DeValois et al., 1982a])
	mean/med: $3.7/2.8$	mean: ≈ 2.2	
		range: $\approx 0.5 - 8.0$	
complex cells	range: $1.8 - 7.8$	bulk $\approx 2.0 - 5.6$	
	mean/med: $3.9/3.2$	mean: 3.2	
		range $\approx 0.5 - 8.0$	
	Frequency bandwidth at 50% amplitude (cycles / deg)		
	Model	Cortex	References
simple cells	range: $1.1 - 1.8$	bulk $\approx 1.0 - 1.5$	[DeValois et al., 1982a]
	med: ≈ 1.45	med: ≈ 1.45	
		range $\approx 0.4 - 2.6$	
complex cells	range: $1.5 - 2.0$	bulk $\approx 1.0 - 2.0$	
	med: 1.6	med: 1.6	
		range $\approx 0.4 - 2.6$	
	Frequency bandwidth at 71% amplitude (index)		
	Model	Cortex	References
simple cells	range: $44 - 58$	bulk $\approx 40 - 70$	[Schiller et al., 1976d]
	med: 55		
complex cells	range $40 - 50$	bulk $\approx 40 - 60$	
	med. 48		
	Orientation bandwidth at 50% amplitude (octaves)		
	Model	Cortex	References
simple cells	range: $38^o - 49^o$	—	[DeValois et al., 1982b]
	med: 44^o		
complex cells	range: $27^o - 33^o$	bulk $\approx 20^o - 90^o$	
	med: 43^o	med: 44^o	
	Orientation bandwidth at 71% amplitude (octaves)		
	Model	Cortex	References
simple cells	range: $27^o - 33^o$	bulk $\approx 20^o - 70^o$	[Schiller et al., 1976c]
	med: 30^o		
complex cells	range: $27^o - 33^o$	bulk $\approx 20^o - 90^o$	
	med: 31^o		

Table 3.1: Summary of the tuning properties of the S_1 and C_1 units *vs.* parafoveal simple and complex cells from monkey primary visual cortex. Model units were probed with the same stimuli as the corresponding studies in cortex (*e.g.*, gratings to assess the orientation bandwidth at 50% amplitude as in [DeValois et al., 1982b] and edges as well as bars of optimal dimensions size to assess the orientation tuning at 71% amplitude as in [Schiller et al., 1976c]. '—' indicates a value we discarded because it appears anomalous and inconsistent with the rest of the literature: as reported in [DeValois et al., 1982b], the median orientation tuning bandwidth of parafoveal complex cells (34^o) would be less than that of both foveal simple (42^o) and complex (45^o) cells.

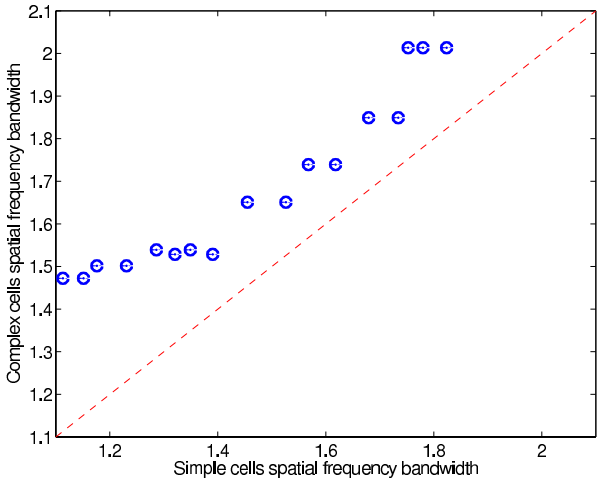


Figure 3-1: Spatial frequency bandwidth of the C_1 *vs.* S_1 units. In the model, we observer an increase of about 20% in the spatial frequency bandwidth of units from the S_1 to the C_1 stage, consistent with parafoveal cortical cells [Schiller et al., 1976d; DeValois et al., 1982a].

A.2 Results

A summary of all the tuning properties of the model units and corresponding primate cortical cells is provided in Table 3.1. Details on how the parameters for the model units were selected as well as all parameter values can be found in Appendix A. Model units seem to capture well the tuning properties of the bulk of parafoveal cells. In addition, consistent with physiology [DeValois et al., 1982a; Schiller et al., 1976d], the population of model units exhibits the following trends:

1. A positive correlation between the size of the receptive fields and the frequency bandwidths;

2. A negative correlation between the size of the receptive fields and the peak frequencies

3. A broadening in the frequency bandwidth from S_1 to C_1 units ($\approx 20\%$, see Fig. 3-1 for illustration).

In Appendix A we describe how S_1 and C_1 parameters were adjusted so that the corresponding units would match the tuning properties of cortical parafoveal cells. The reader can refer to [Serre and Riesenhuber, 2004] for a comparison between the new model parameters and the parameters of the original model [Riesenhuber and Poggio, 1999a].

B V4 and the Model

Here we compare the tuning properties of model C_2 units to the tuning properties of cells in V4. As described in Chapter 2, the tuning of the S_2 and C_2 units is learned from natural images during an unsupervised learning stage. During this developmental-like learning stage, model units become tuned to image features (patches of natural images) that appear with high probability in natural images. Here we take a closer look at the detailed tuning properties of the resulting units and analyze their selectivity to standard stimuli. We show that C_2 units that are learned from natural images exhibit tuning properties that agree with experimental data from V4. First we analyze the response of model C_2 units to boundary conformations [Pasupathy and Connor, 2001] (see Section B.1); second we look at the interaction of two-bar stimuli when presented within the receptive fields of S_2 units [Reynolds et al., 1999] (in the absence of attention, see Section B.2).

B.1 Tuning to Boundary-Conformation Stimuli

To try to get a more quantitative description of V4 cells, Pasupathy & Connor considered a parametrized space of moderately complex 2D shapes (see Fig. 3-2) to probe V4 neurons. The stimulus dataset was generated by systematically combining convex and concave boundary elements to produce simple closed shapes with shared boundary components. With this parameterized stimulus dataset they quantified the responses of prescreened V4 cells (responsive to complex stimuli) and look at which ones, from a set of different tuning spaces, best explained the data. They compared three different tuning domains:

1. Tuning for boundary conformation, *i.e.*, characterizing the neural response in a *curvature* $\times$ *angular position* tuning space, where *curvature* is defined as the rate of change in tangent angle with respect to contour length (see [Pasupathy and Connor, 2001] for details) and *angular position* is measured with respect to the object center of mass;

2. Tuning for linear edge orientation;

3. Tuning for axial orientation (where axial denotes the axis of greatest elongation).

Pasupathy & Connor found that the boundary conformation space best characterized the set of 109 V4 responses they recorded from. From these results, they suggested a part-

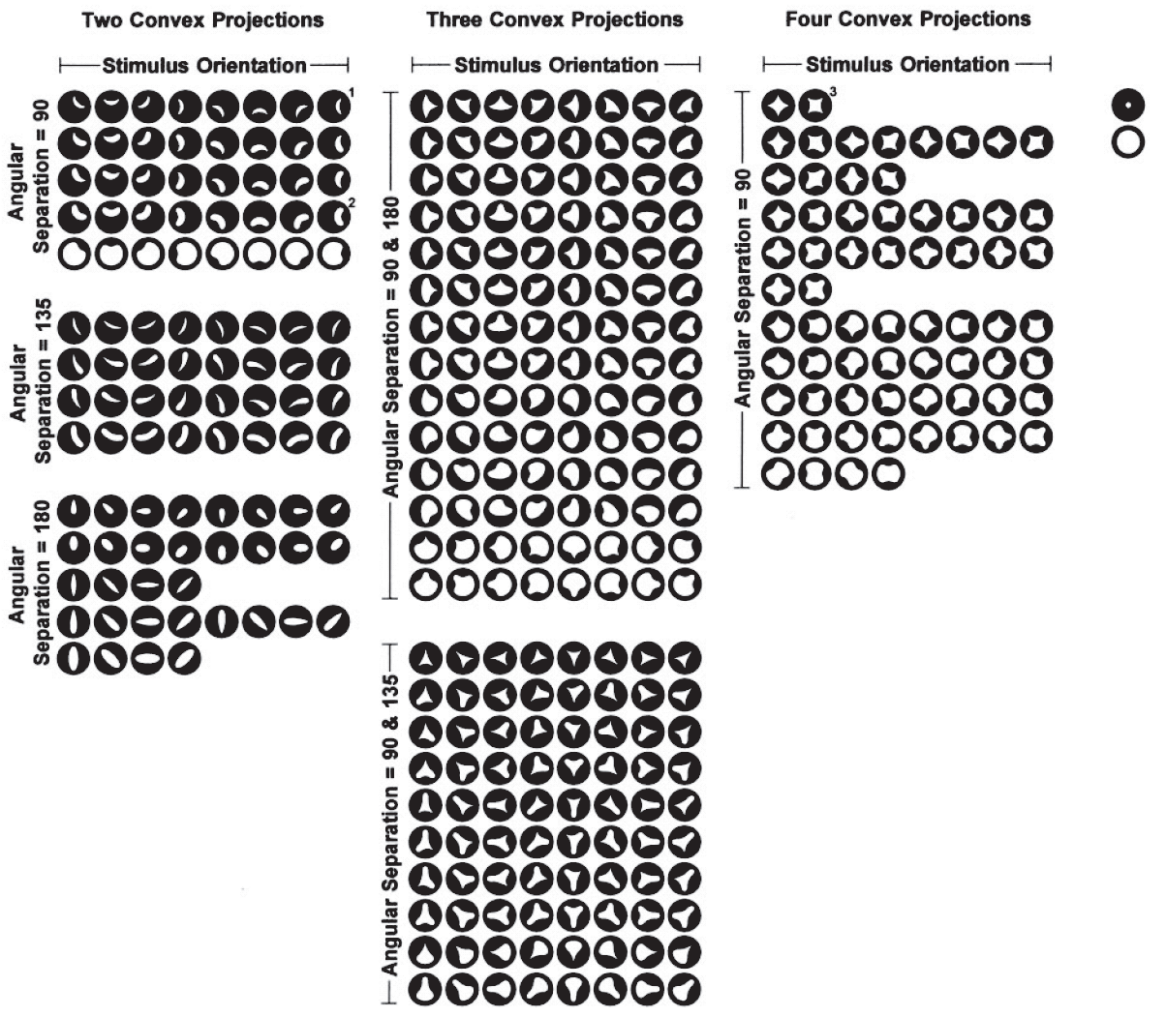


Figure 3-2: The stimulus dataset used to measure the tuning of cells to boundary conformations (see [Pasupathy and Connor, 2001] for details).

based representation of complex shapes in V4, where the parts are boundary patterns defined by curvature and position relative to the rest of the object.

To look at the plausibility of the architecture of Fig. 2-1 and the associated learning rule we performed a similar "recording" experiment on model units. Using the same stimulus set as in [Pasupathy and Connor, 2001] (see Fig. 3-2) we recorded from 109 model units in the C_2 layer that were pre-screened for their selectivity to complex shapes. As in [Pasupathy and Connor, 2001] different Gaussian tuning functions (*i.e.*, axial orientation, edge orientation and boundary conformation) were fit to the response of the recorded units.

Results

Fig. 3-3 shows a comparison between one V4 neuron (from Fig. 4A in [Pasupathy and Connor, 2001]) and one of the C_2 unit we recorded from. Both seem to be tuned to concave curvature at the right and exhibit qualitatively similar patterns of responses to the different shapes. The model unit was picked from the population of 109 model C_2 units under study. Both units exhibit very similar pattern of responses (overall correlation $r = 0.78$). The fit between the model unit and the V4 neuron is quiet remarkable given that there was no fitting procedure involved here for learning the weights of the model unit: The unit was simply selected from the small population of 109 model units that we recorded from; its tuning (or preferred stimulus) was learned from natural images. The organization of the receptive field of the C_2 units was dictated by the theory described in Chapter 2: The shape selectivity of the C_2 units is inherited from its afferent S_2 units that are all tuned to the same preferred stimulus but centered at slightly different positions and scales (thus providing tolerance to shift and size). Each S_2 afferent itself receives its inputs from oriented V1-like complex cells (see inset of Fig. 3-3(b) for an illustration of the organization of one S_2 unit).

Fig. 3-4 displays the results of the simulated experimental methodology from [Pasupathy and Connor, 2001] performed on the overall population of 109 model units. Each model unit has learned the pattern of input from a natural image patch. Goodness-of-fit was assessed by calculating the coefficient of correlation between neural responses and responses predicted by the tuning functions (see Fig. 3-4). The resulting population of model units exhibits tuning that is best explained by the tuning in the boundary conformation space than in the edge orientation space or axial orientation space, a characteristic of V4 neural population as reported by [Pasupathy and Connor, 2001]. The median values of the correlation coefficients of the model units and V4 neurons for different tuning functions are summarized in Table 3.2. The V4 data is courtesy of Pasupathy & Connor. [2]

Methods

Below we describe the methodology used to assess the selectivity of units for boundary conformations ([see Pasupathy and Connor, 2001] for details).

Stimulus Set: The stimulus set used in [Pasupathy and Connor, 2001] contains 366 stimuli created by systematically combining convex and concave boundary elements. The stimulus dataset (see Fig. 3-2) was reproduced using code kindly supplied by Anitha Pasupa-

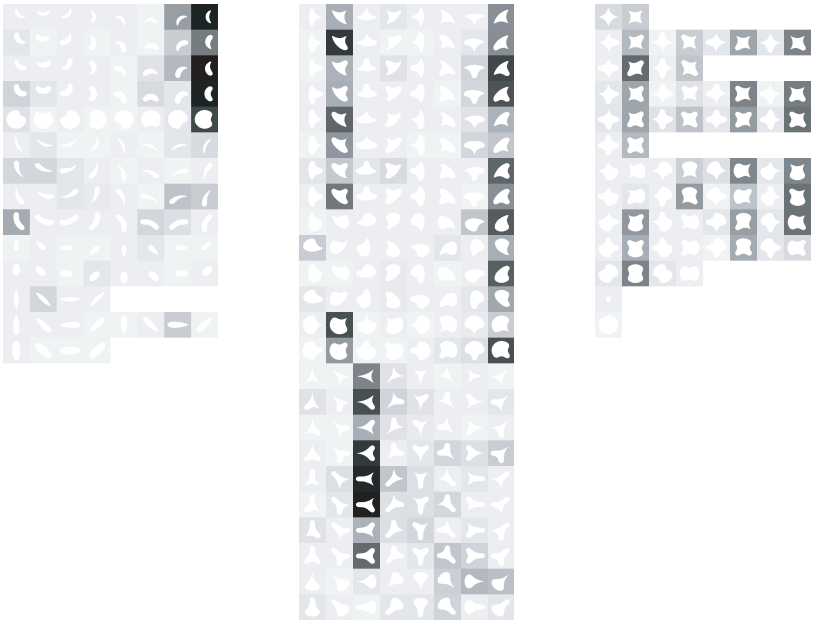
Tuning functions	Model units	V4 neurons
2-D boundary conformation	0.38	0.41
4-D boundary conformation	0.47	0.46
2-Gaussian boundary conformation	0.50	0.46
Edge orientation	0.11	0.15
Edge orientation + contrast polarity	0.18	0.21
2-D axial orientation $\times$ elongation tuning functions	0.28	0.18
3-D axial orientation $\times$ length $\times$ width tuning functions	0.32	0.28

Table 3.2: Goodness of fit (median value of the correlation coefficients across all cells) of different tuning functions for model units and V4 neurons.

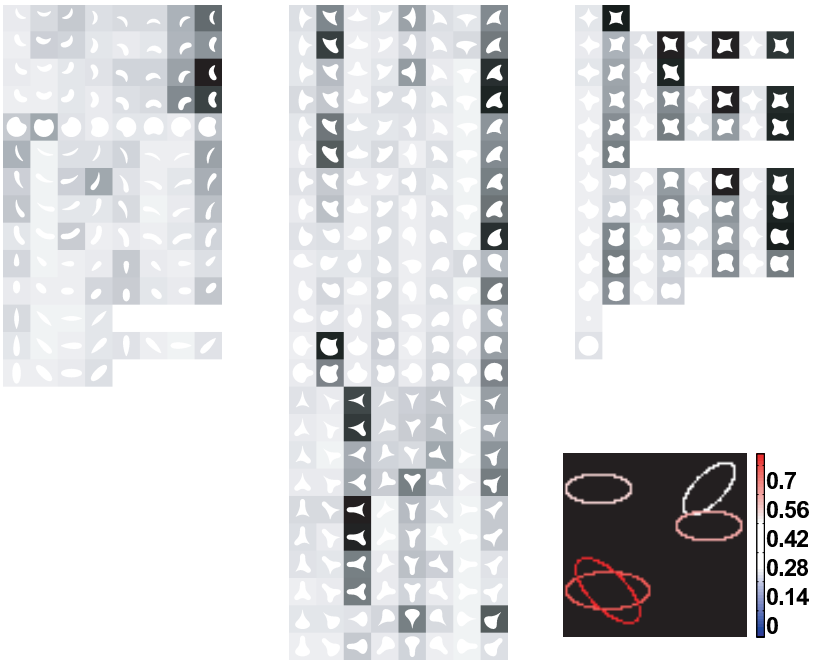
thy. Each stimulus is represented by a white icon drawn within a black circle representing the unit receptive field. Details on the construction of the stimulus set and data analysis can be find in [Pasupathy and Connor, 2001].

Tuning Spaces Each boundary element was characterized by four numbers (curvature, orientation, angular position, and radial position) and could be considered a point in a multidimensional space such that each shape is a collection of such points. The tuning of model units was characterized using the same shape space analysis as used by Pasupathy & Connor. Multi-dimensional Gaussian functions were fit for each model unit in a shape space based on the stimuli. The multi-dimensional functions used to characterize model responses are: (a) 2-D boundary conformation, (b) 4-D boundary conformation, (c) 2-Gaussian boundary conformation, (d) edge orientation, (e) edge orientation + contrast polarity, (f) 2-D axial orientation $\times$ elongation tuning functions and (g) 3-D axial orientation $\times$ length $\times$ width tuning functions.

The 2-D boundary conformation domain represents the contour elements of each stimuli in a curvature $\times$ angular position space. The 4-D boundary conformation domain contains, in addition to the same curvature $\times$ angular position space as the 2-D boundary conformation space, two adjacent curvature dimensions (*i.e.*, the central curvature is augmented by the curvatures of the contour segments that are counterclockwise and clockwise adjacent). An edge orientation shape space analysis was also used to determine if responses were selective to flat contour segments at specific orientations. For this space each contour segment of a stimulus was parameterized by the angle between the tangent line and the horizontal. As in [Pasupathy and Connor, 2001], for each model unit, we characterized its tuning in shape space by deriving multi-dimensional Gaussian functions based on neural responses (see [Pasupathy and Connor, 2001] for details).

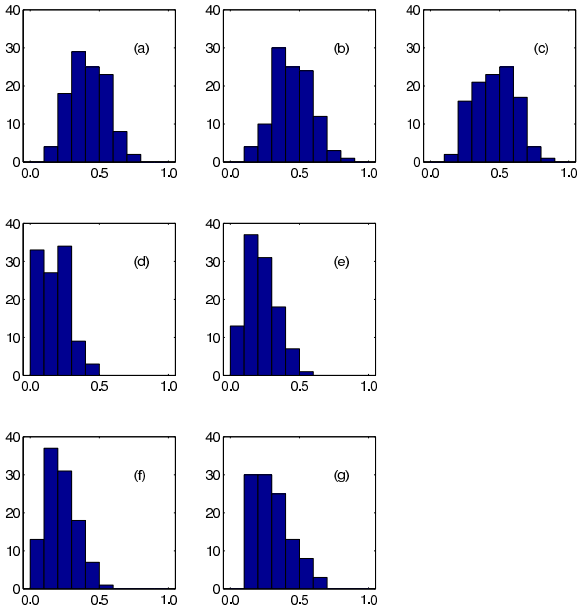


(a) One V4 Neuron

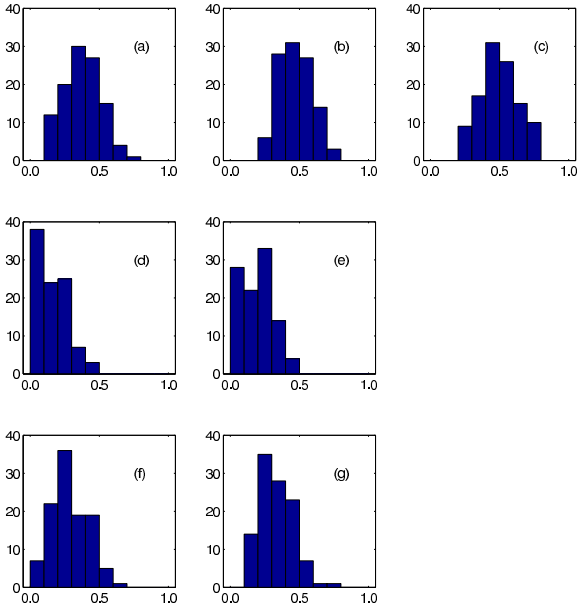


(b) One model C_2 unit

Figure 3-3: A comparison between the response of a single V4 neuron (corresponding to Fig. 4A in [Pasupathy and Connor, 2001]) **(a)** and a single model C_2 unit **(b)** over the boundary conformation stimulus set. The response magnitude to each stimulus is indicated by the gray level of the stimulus background. The darker the shading the stronger the response. The model unit was picked from the population of 109 model C_2 units under study. Both units exhibit very similar pattern of responses (overall correlation $r = 0.78$). The fit between the model unit and the V4 neuron is quiet remarkable given that there was no fitting procedure involved here for learning the weights of the model unit: The unit was simply selected from a small population of 109 model units learned from natural images and selected at random. The inset on the lower right end of the figure at the bottom describes the corresponding receptive field organization of the C_2 unit. Each oriented ellipse characterizes one subfield at matching orientation. Color encodes for the strength of the connection between the subfield and the unit.



(a) Population of V4 Neurons



(b) Population of model C_2 units

Figure 3-4: Model C_2 units **(b)** exhibit tuning properties comparable with V4 data **(a)** from [Pasupathy and Connor, 2001]. In each figure, each one of the seven panels displays the population histogram of the correlation coefficient (goodness of fit) for a different tuning function (see text): *i.e., boundary conformation* (top row) with 2-D boundary conformation (a), 4-D boundary conformation (b) and 2-Gaussian boundary conformation (c), edge orientation (middle row) with edge orientation (d) and edge orientation and contrast polarity (e) as well as axial orientation (bottom row) with 2-D axial orientation $\times$ elongation tuning functions (f) and 3-D axial orientation $\times$ length $\times$ width tuning functions (g). V4 neurons characteristically show higher correlation coefficients for boundary conformation tuning functions (a, b, c) than for edge orientation or axial orientation tuning functions (d, e, f, g) and so do the model units.

B.2 Tuning to Two Bar-Stimuli

We here look at the response of model C_2 units to single and two-bar stimuli presentations as [Reynolds et al., 1999]. [Reynolds et al., 1999] recorded the response of individual V4 neurons in two experimental conditions, *i.e.*, when the monkey was attending either outside or inside the receptive field of neurons. The feedforward model described in this thesis does not yet account for attentional mechanisms. Therefore in the following we only look at the condition for which the monkey was attending away from the receptive field such that attention did not affect the response of the recorded neuron. In this condition, Reynolds *et al*. found that the addition of a second stimulus presented within the receptive field of a V4 neuron (see Fig. 3-5) causes the response of the neuron to move toward the response of the second stimulus alone.

B.3 Results

We performed a similar experiment while "recording" from a population of model C_2 units. As in [Reynolds et al., 1999], we presented model units with both a reference stimulus and a probe stimulus (each an oriented bar) either at the same time or individually. The responses were then analyzed using a similar methodology as in the experiment (see Section B.3).

The results of this experiment are shown in Fig. 3-6. The experimental findings are reproduced on the left, and the model results are on the right. Interestingly, such interference effects (*i.e.*, presenting a preferred and a non-preferred stimulus together produces a neural response that falls between the neural responses to the two stimuli individually, sometimes close to an average, in the absence of attention) may be occurring in other cortical areas such as IT, see [Serre et al., 2005a]. The model accounts for such "clutter effect" regardless of any particular cortical area, using the same principle operations for selectivity and invariance appearing across different layers. In fact, the biased competition model devised to explain the results of [Reynolds et al., 1999] is closely related to Eq. 1.3 in our model. Since normal vision operates with many objects appearing within the same receptive fields and embedded in complex textures (unlike the artificial experimental setups), understanding the behavior of neurons under such clutter condition is important and warrants more experiments.

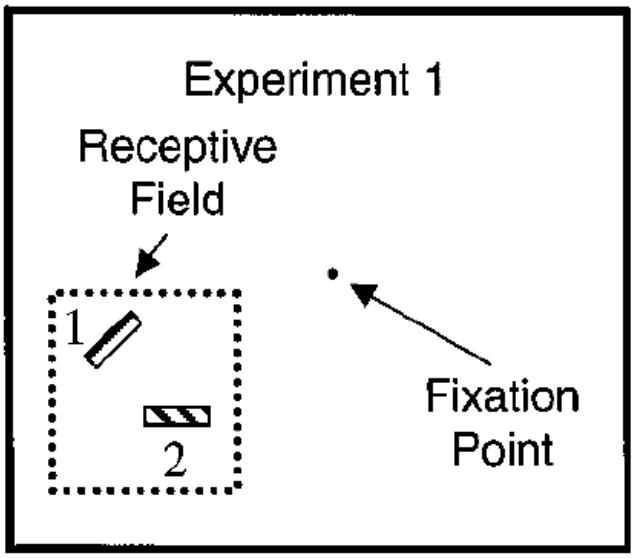
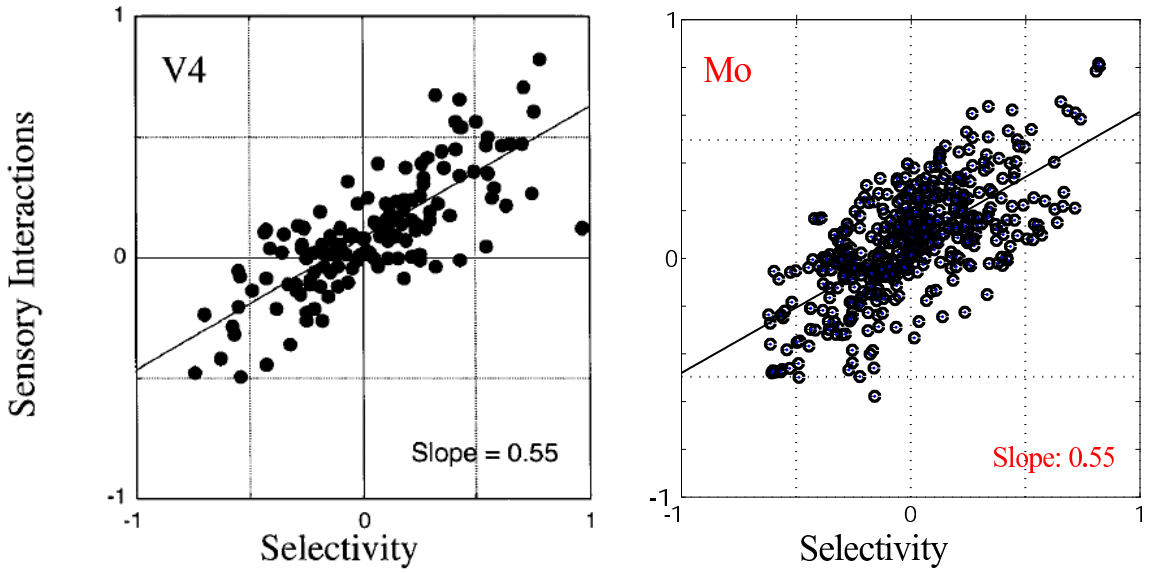


Figure 3-5: The two-bar stimulus experiment by Reynolds *et al.* (modified from [Reynolds et al., 1999]). The monkey here is attending away from the receptive field of the neuron being recorded. Reynolds *et al.* found that the addition of a second stimulus presented within the receptive field of a V4 neuron causes the response of the neuron to move toward the response of the second stimulus alone.

Methods

As in [Reynolds et al., 1999] we used 16 stimuli. However, because the model does not yet include color sensitive cells, instead of using 16 stimuli composed of all combinations of four oriented bars (0^o, 45^o, 90^o, and 135^o) presented in four colors (red, blue, green, and yellow), we presented 16 gray-scale oriented bars. Thus an equal number of measurements was performed on model units and V4 neurons.

As in [Reynolds et al., 1999], stimuli could appear at one of two possible locations within the receptive field of units: By definition, the stimulus that appears at position one (see Fig. 3-5) is designated as the reference stimulus (chosen from the set of 16 possible stimuli). As in [Reynolds et al., 1999], the reference stimulus was chosen sometimes to be the preferred stimulus for the unit, sometimes the weakest and sometimes an intermediate stimulus. The stimulus that appears at position two, designated the probe stimulus, was selected at random from the same set of 16 possible stimuli as in [Reynolds et al., 1999]. While the identity of the probe (if presented) varied for each trial, the identity of the reference was fixed throughout the entire "recording" session. On any given trial, we tested three conditions:


V4
Slope = 0.55
Sensory Interactions
Selectivity
1
0
-1
-1
0
1
Mo
Slope: 0.55
Selectivity
1
0
-1
-1
0
1