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

CSAIL

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

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

List of Figures

List of Tables

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-

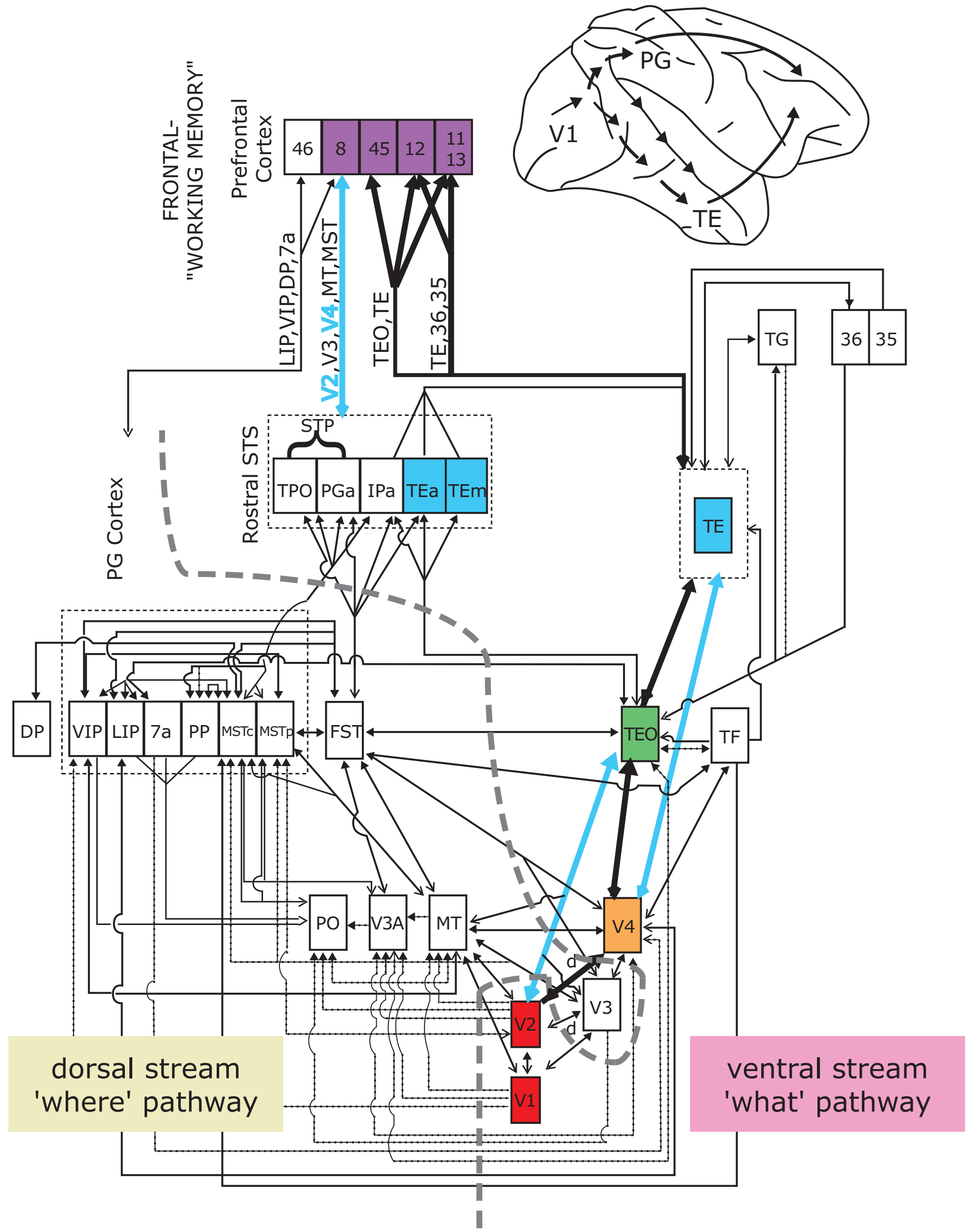


Figure 1-1: The ventral stream of visual cortex and object recognition. Modified from Ungerleider & Van Essen [Gross, 1998].

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].

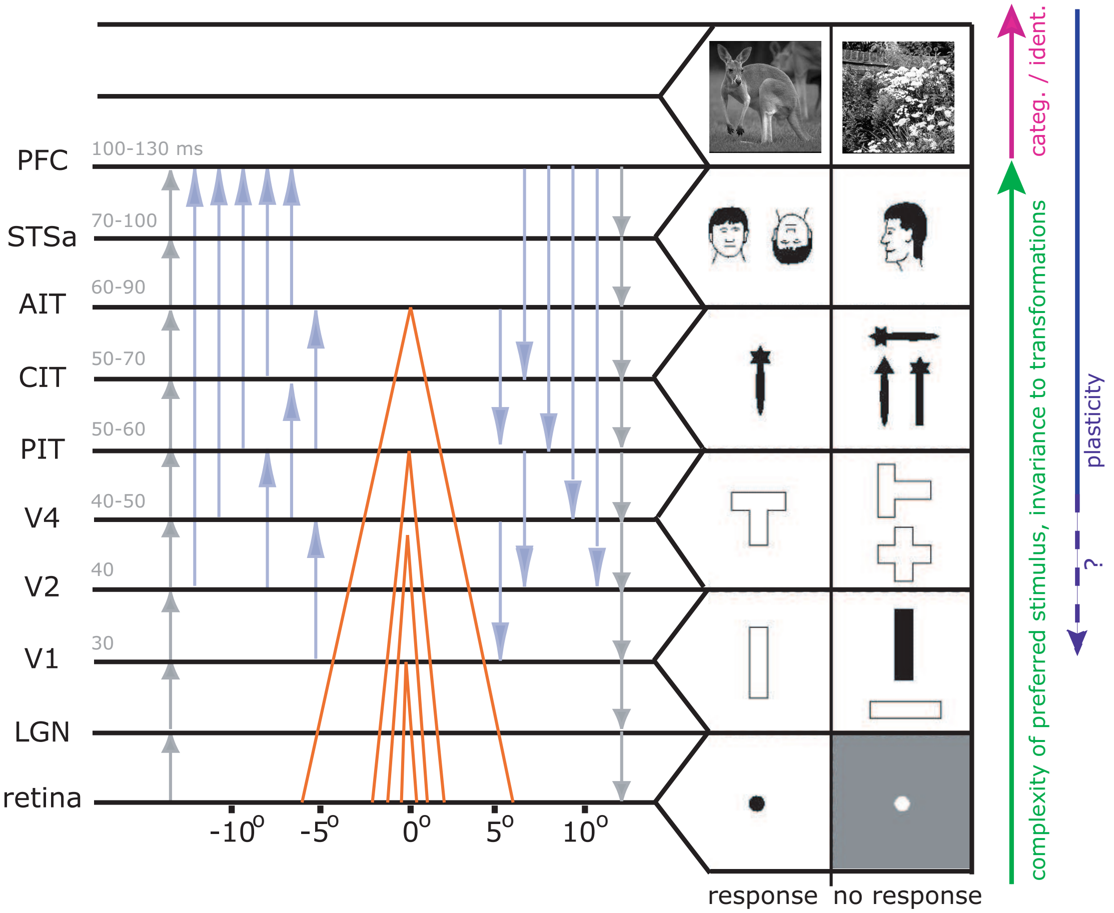


Figure 1-2: The organization of visual cortex based on a core of knowledge that has been accumulated over the past 30 years. The figure is modified from [Oram and Perrett, 1994] mostly to include the likely involvement of prefrontal cortex during recognition tasks by setting task-specific circuits to 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].

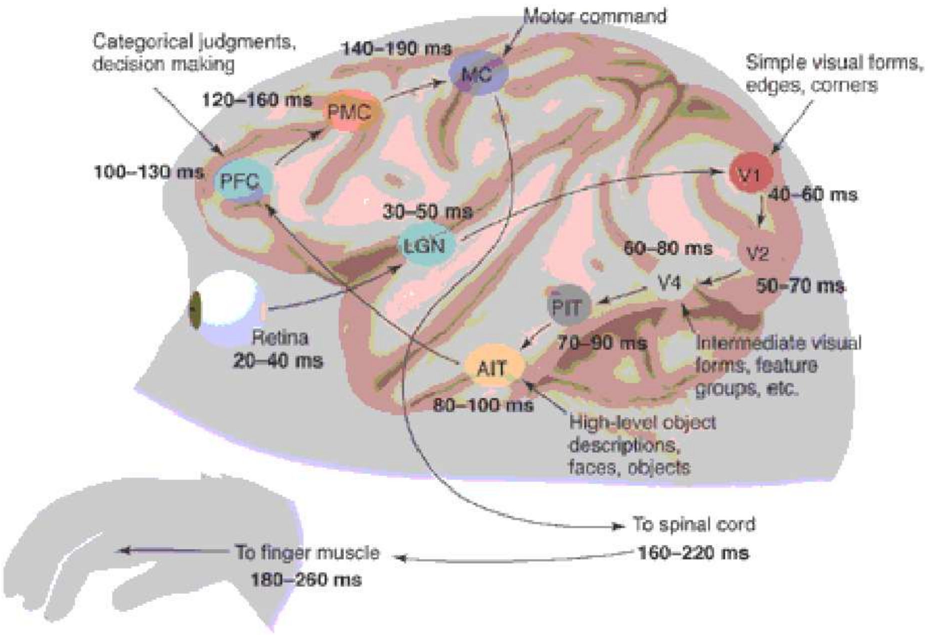


Figure 1-3: The feedforward circuits involved in rapid categorization tasks. Numbers for each cortical stage corresponds to the shortest latencies observed and the more typical mean latencies [Nowak and Bullier, 1997; Thorpe and Fabre-Thorpe, 2001]. Modified from [Thorpe and Fabre-Thorpe, 2001].

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).

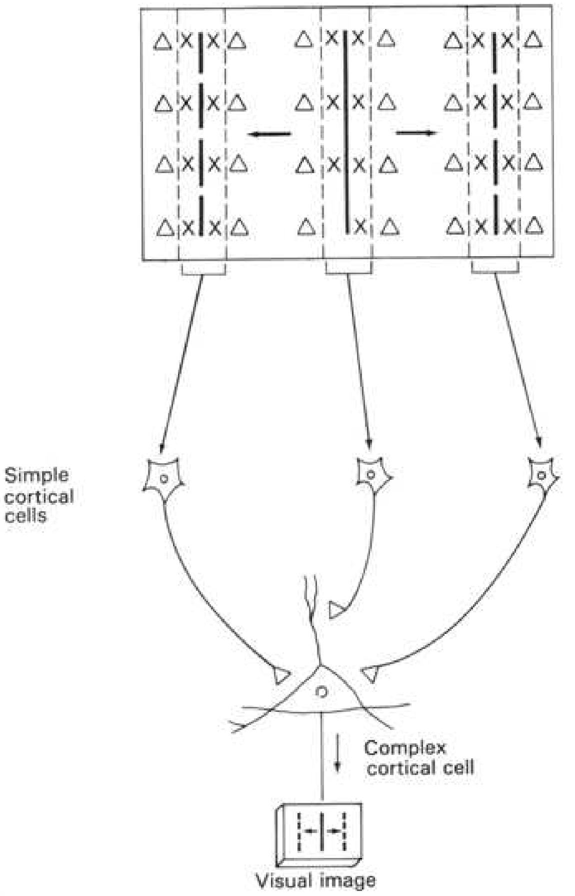


Figure 1-4: The Hubel & Wiesel hierarchical model for building complex cells from simple cells. Reproduced from [Hubel and Wiesel, 1959].

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-

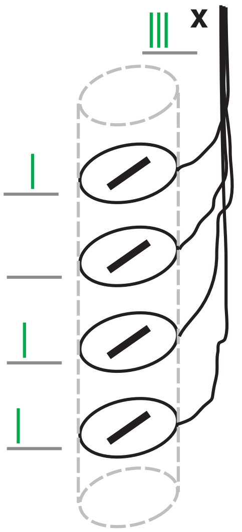


Figure 1-5: Computational modules in cortex: The basic processing unit in models of object recognition in cortex (*e.g.*, the nodes of Fig. 2-1) may correspond to *computational modules* in cortex rather than single neurons. Modules would be composed of *equivalent* cells with identical parameters and 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

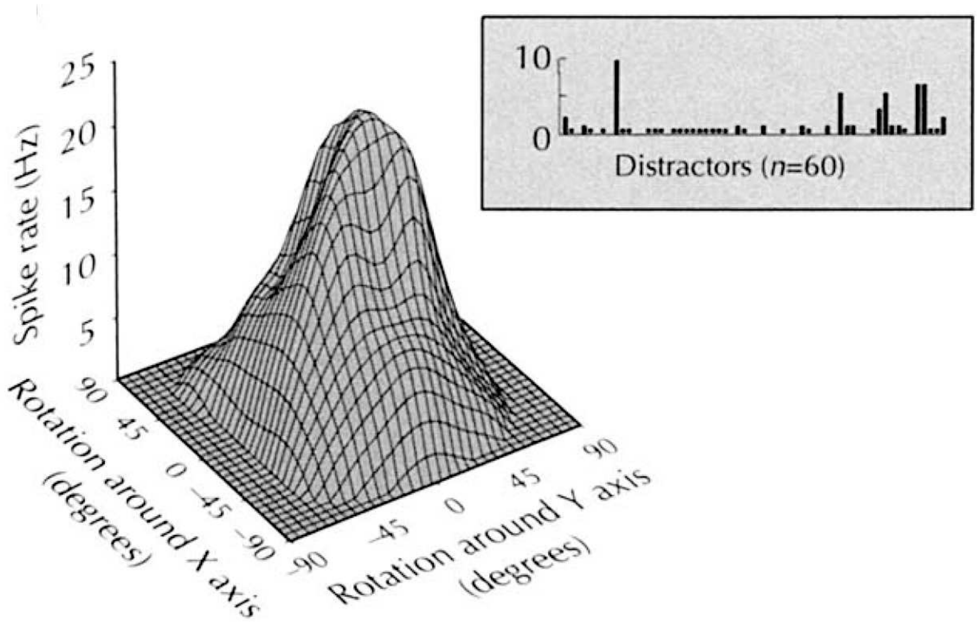


Figure 1-6: A typical bell-shaped TUNING from one cell in AIT. Illustrated here is the tuning of a particular cell to a specific view of a paperclip presented during training [Logothetis et al., 1995]. As the stimulus presented rotates away from the tuned view (along either of the two axes), the response 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 cortex. 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.

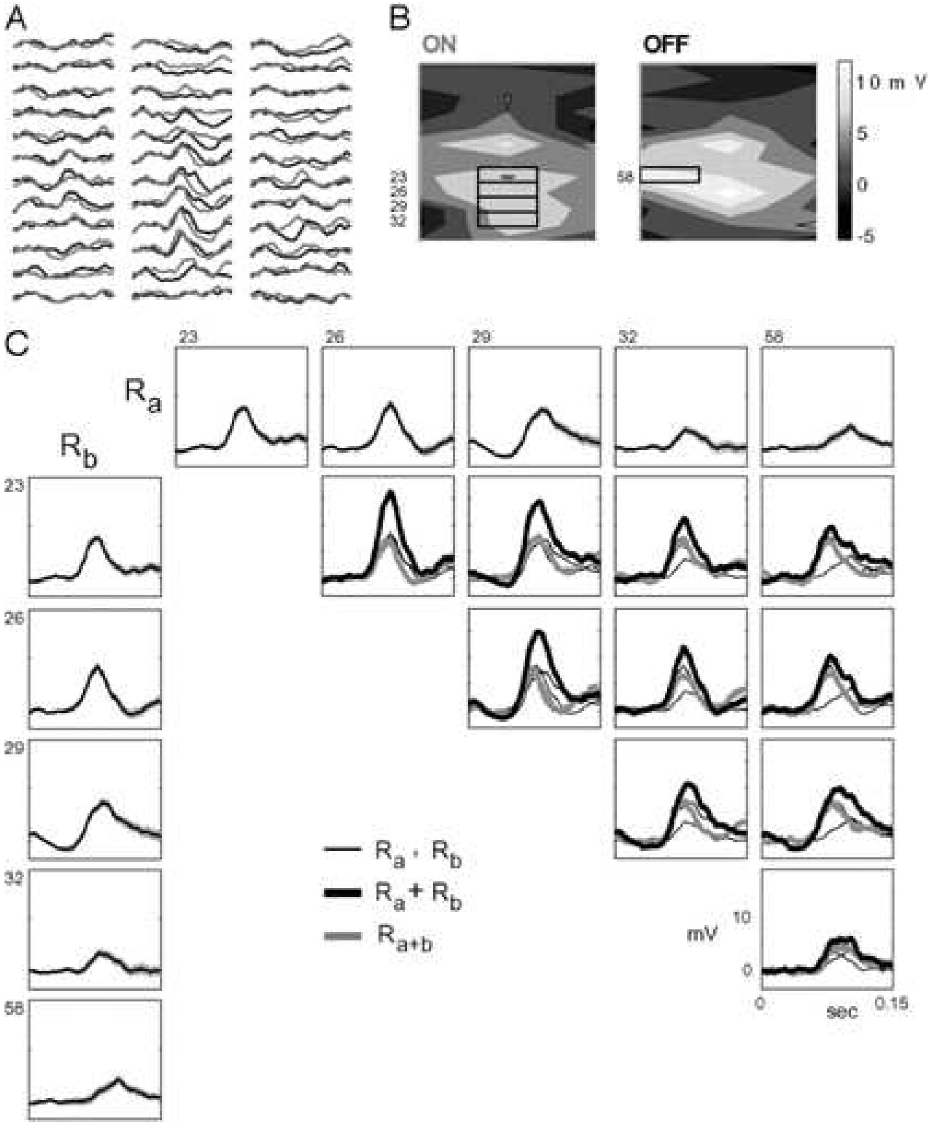


Figure 1-7: MAX operation from a complex cell in area 17 of the cat. Illustrated is the response of a complex cell to the simultaneous presentation of two bars (see [Lampl et al., 2004]) for details). A: average membrane potential measured from the response of the cell to bars of the optimal orientation. 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:

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

where $k << 1$ is a constant to avoid zero-divisions and p, q and r represent the static nonlinearities 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.*, $||\mathbf{x}||^{\mathbf{2}} = ||\mathbf{w}||^{\mathbf{2}}$ =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 $\mathbf{p} \gtrapprox \mathbf{q+1}$ $(\mathbf{w_j \approx 1})$, 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]

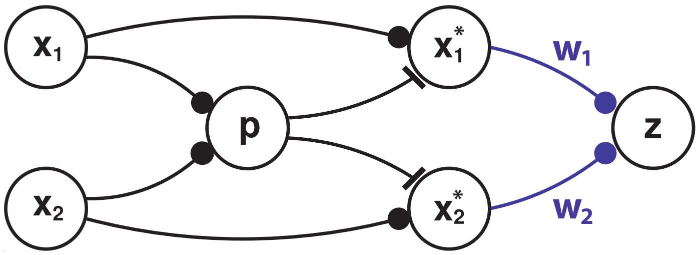


Figure 1-8: A possible cortical circuit for TUNING and MAX operations proposed by Knoblich & Poggio. The nodes in the circuit correspond to computational modules composed of equivalent units, see Fig. 1-5. Preliminary results suggest that depending on the balance between excitation and inhibition (see 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 recognition 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]

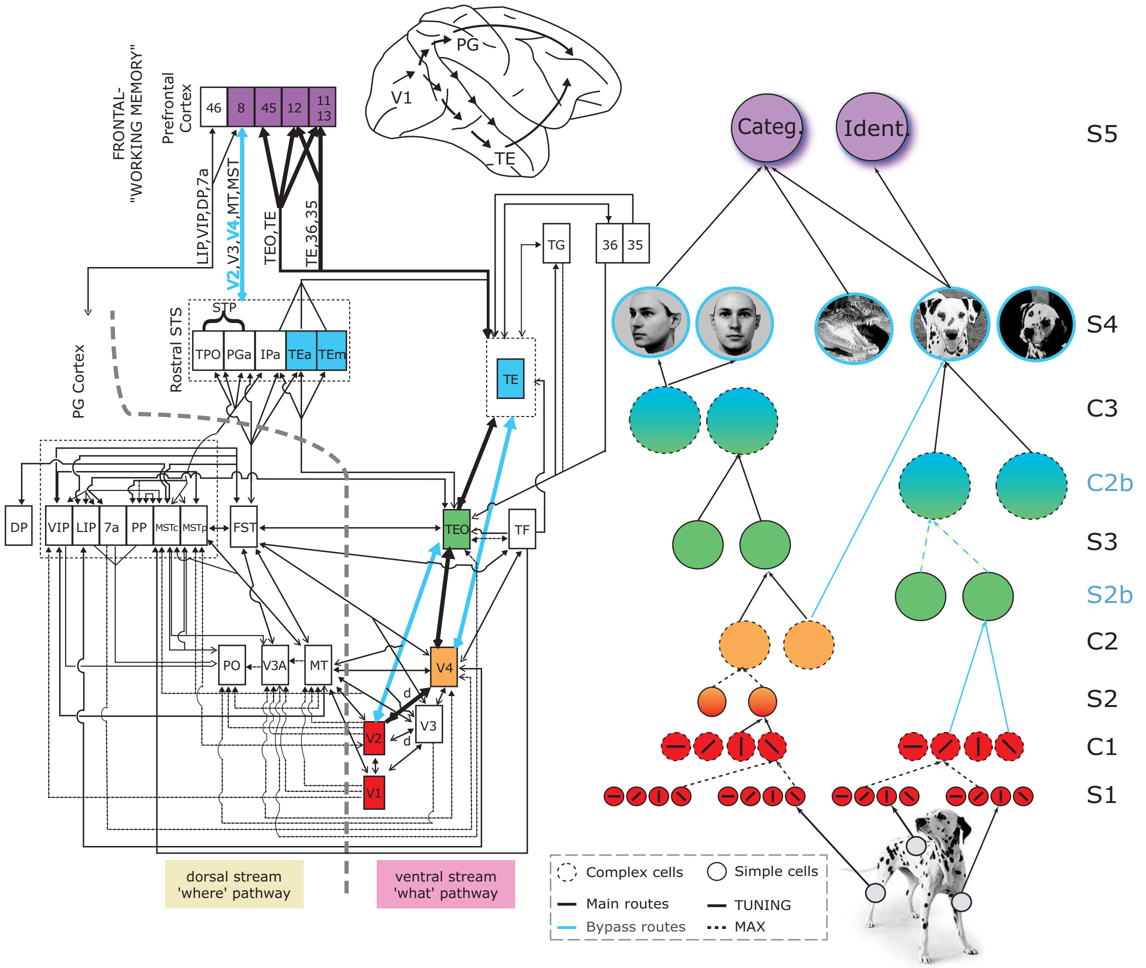


Figure 2-1: Tentative mapping between (right) the functional primitives and layers of the model and (left) cortical stages in the primate visual system (modified from Van Essen & Ungerleider [Gross, 1998]). The correspondences are illustrated with colors (see text). Stages of *simple* (*S*) units (plain 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.

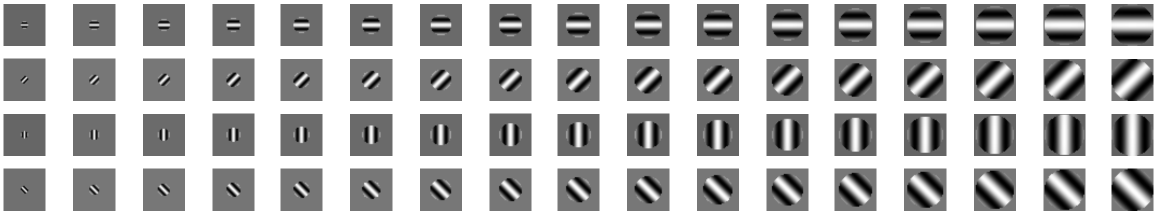


Figure 2-2: Receptive field organization of the S_1 units. There are 136 different types of S_1 units: 2 phases × 4 orientations × 17 sizes (or equivalently peak frequencies). Only units at one phase are shown but the population also includes filters of the opposite phase. Receptive field sizes range 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 × 4 orientations × 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

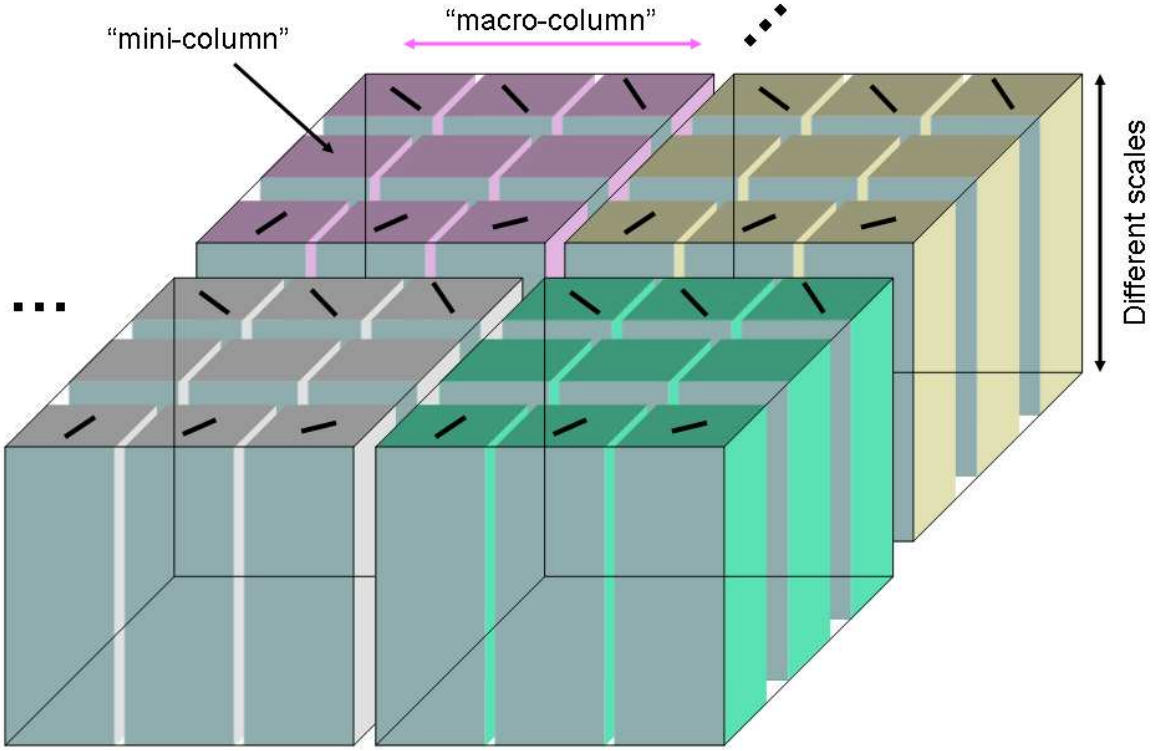


Figure 2-3: Functional "columnar" organization in the model. Each basic *mini-column* contains a set of units all with the same selectivities, *i.e.,* sharing the same weight vector $\mathbf{w}$ (*e.g.,* a bar at a particular orientation at the S_1 level) but different scales (*e.g.,* 17 different scales/peak 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-

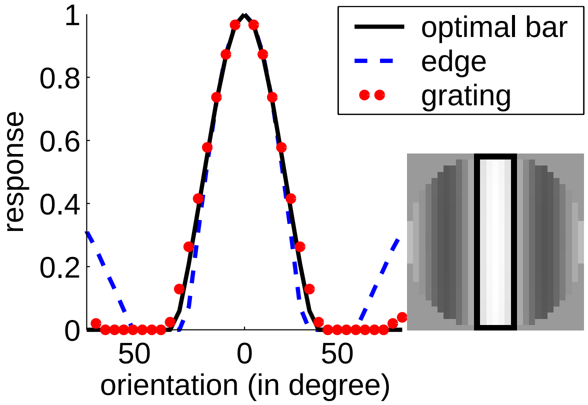


Figure 2-4: One S_1 unit and its corresponding orientation tuning curve obtained with three classical stimuli, *i.e.,* optimal bar (see black rectangle superimposed on the unit receptive field), optimal grating and edge.

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

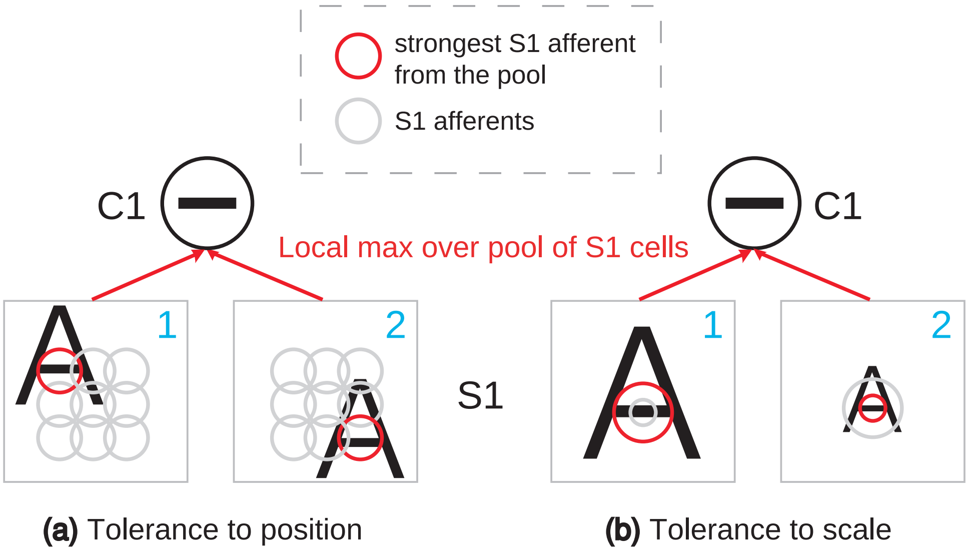

strongest S1 afferent from the pool
S1 afferents
C1
Local max over pool of S1 cells
C1
1
2
S1
1
2
(a) Tolerance to position
(b) Tolerance to scale

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.

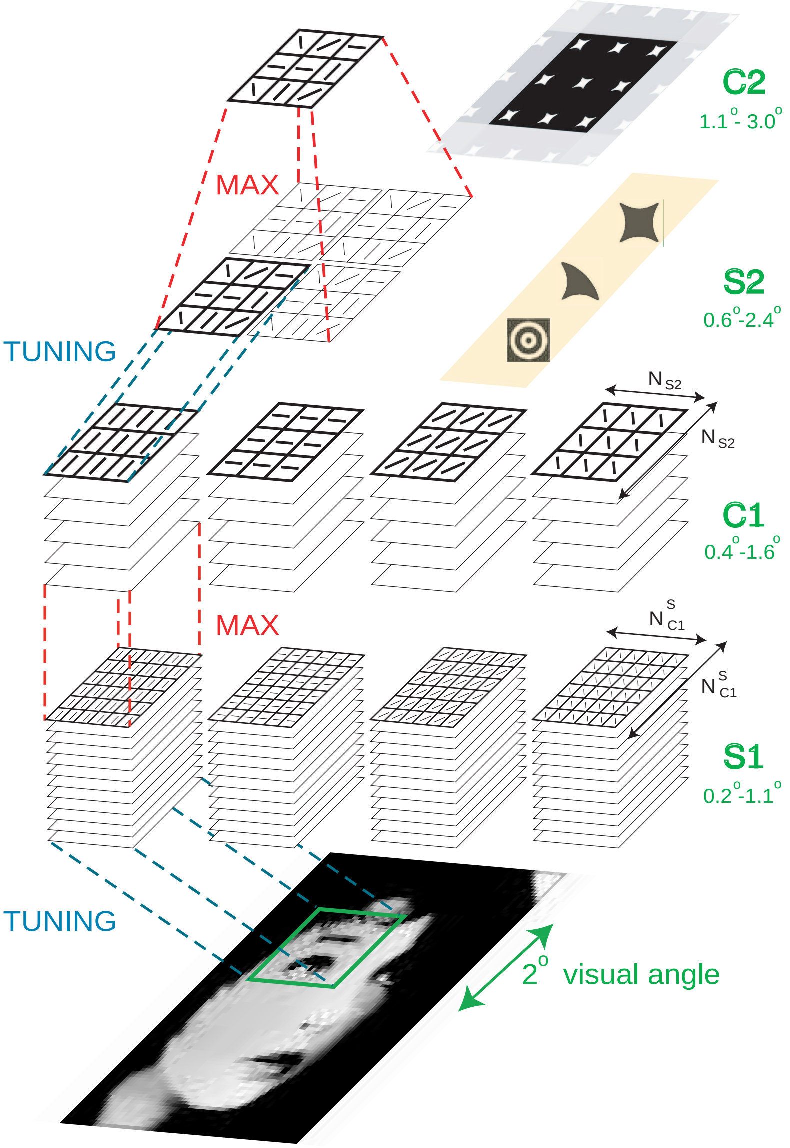


Figure 2-6: Building S_2 and C_2 units. A gray-value input image is first analyzed by functionally organized (see Fig 2-3) S_1 units at all locations. At the next C_1 layer, a local MAX pooling operation is taken over retinotopically organized S_1 units at neighboring positions and scales but with the *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]. Indeed 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 $\mathbf{w}$ (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 $f^o \lesseqgtr \theta$.

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 $f^o(\mathbf{x}) \lesseqgtr \theta$. 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*).

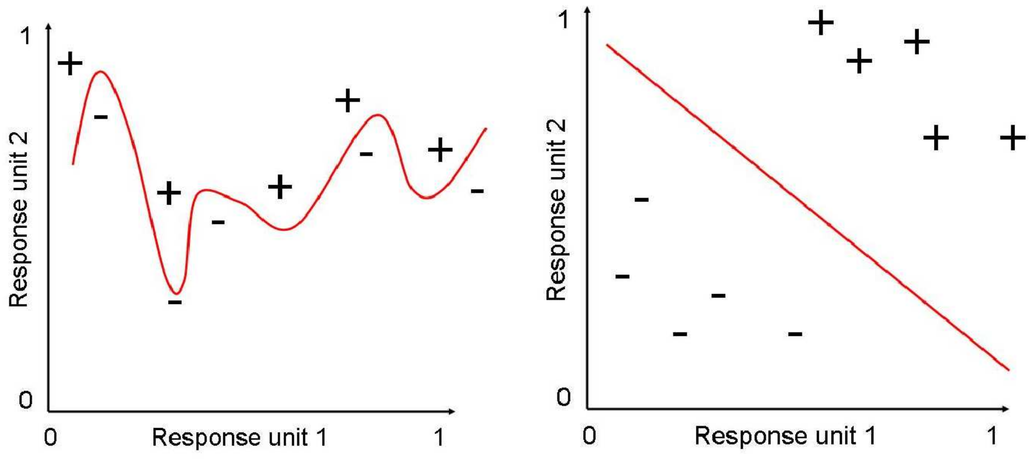
D Discussion

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.

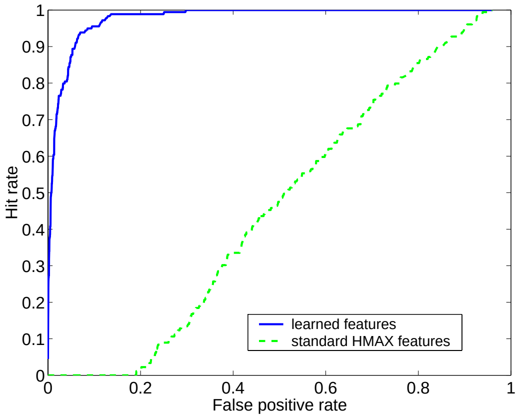


Figure 2-9: Face detection in natural images: Comparison between "learned features" [Serre et al., 2002] and "standard HMAX features" [Riesenhuber and Poggio, 1999a]. The gain in performance after introducing learning in the model is significant. Also note that the "learned feature" version corresponds 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 overfitting of the *object-specific* set that learns both the features and the discrimination function

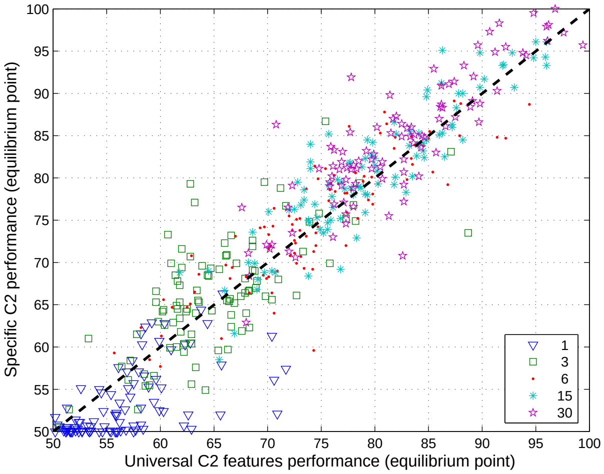


Figure 2-10: *Expert vs. generic* features on the *CalTech-101* object database (see Chapter 4). For each of the 101 object category available for testing and for different number of examples available for training, we compare the performance of a linear classifier that uses a *generic* ("universal", 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.

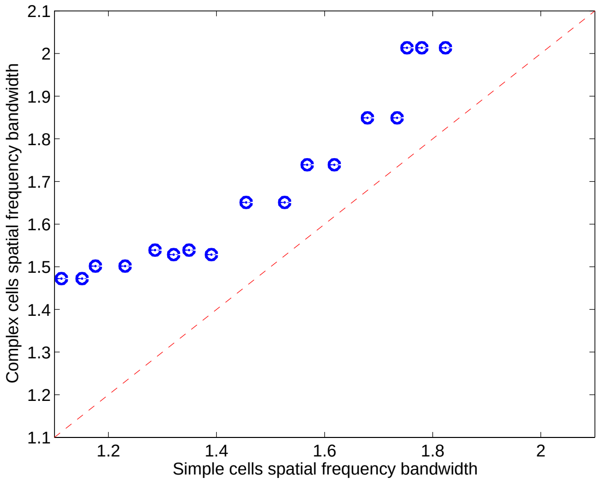


Figure 3-1: Spatial frequency bandwidth of the C_1 *vs.* S_1 units. In the model, we observer an increase of about 20% in the spatial frequency bandwidth of units from the S_1 to the C_1 stage, consistent with parafoveal cortical cells [Schiller et al., 1976d; DeValois et al., 1982a].

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* × *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-

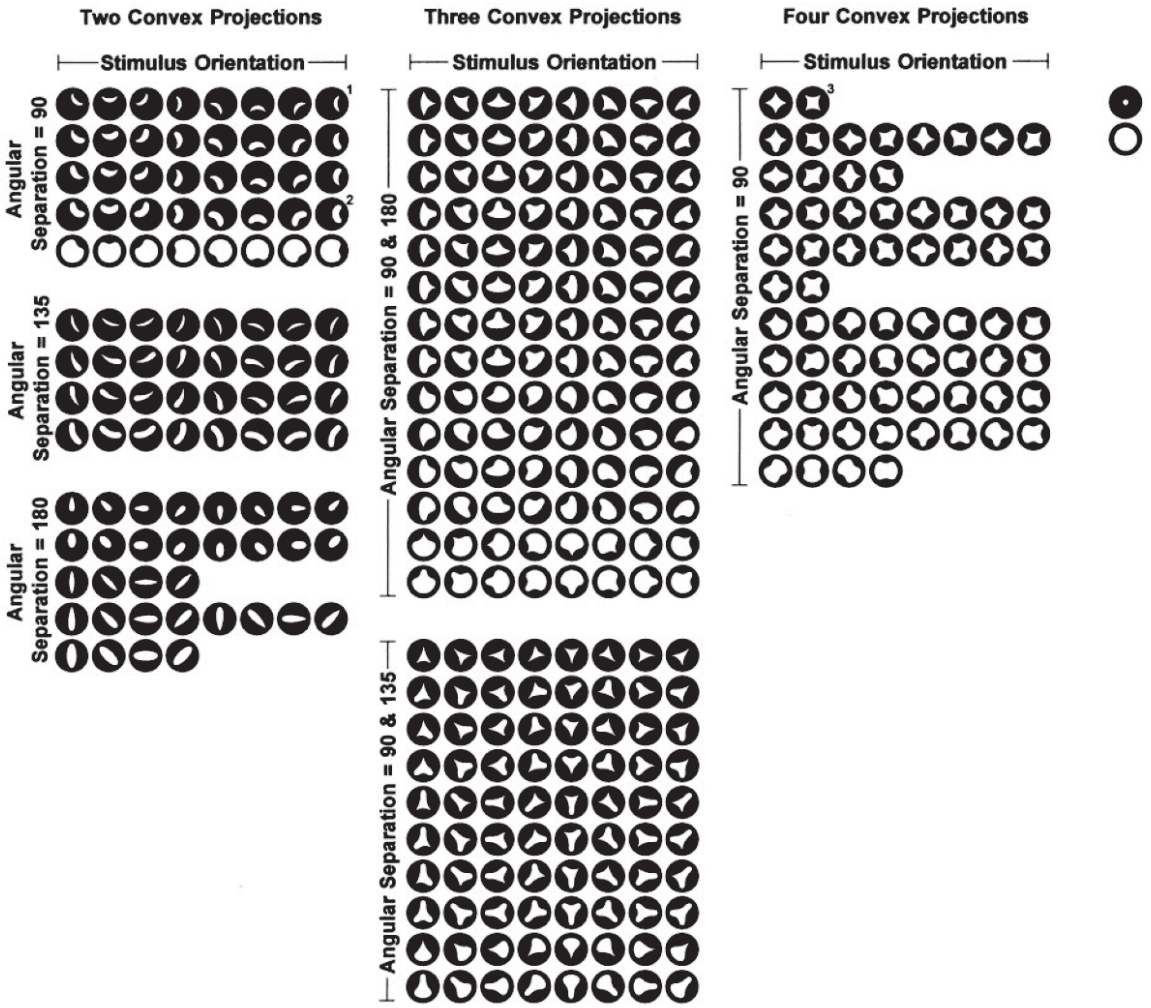


Figure 3-2: The stimulus dataset used to measure the tuning of cells to boundary conformations (see [Pasupathy and Connor, 2001] for details).

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-

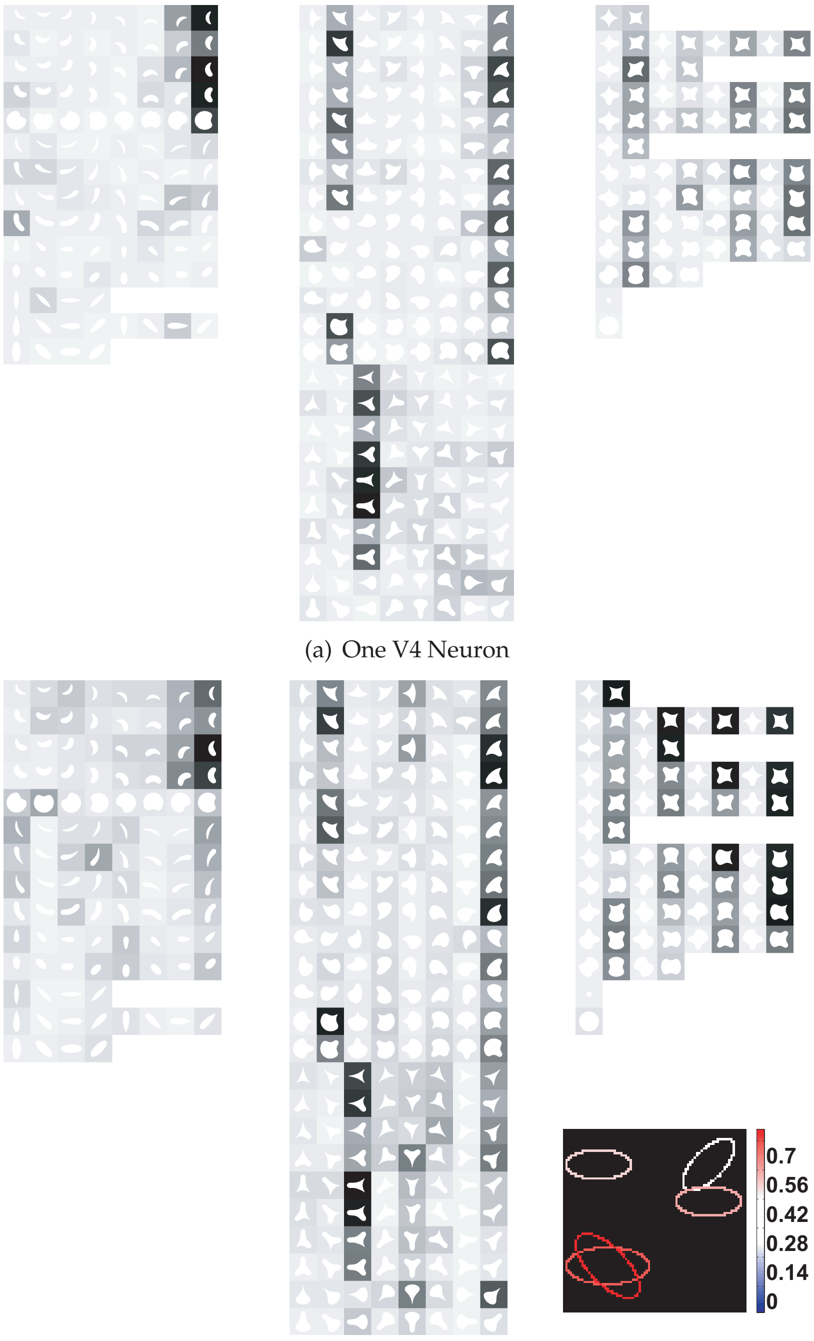
Tuning functions	Model units	V4 neurons
2-D boundary conformation	0.38	0.41
4-D boundary conformation	0.47	0.46
2-Gaussian boundary conformation	0.50	0.46
Edge orientation	0.11	0.15
Edge orientation + contrast polarity	0.18	0.21
2-D axial orientation × elongation tuning functions	0.28	0.18
3-D axial orientation × length × 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 × elongation tuning functions and (g) 3-D axial orientation × length × width tuning functions.

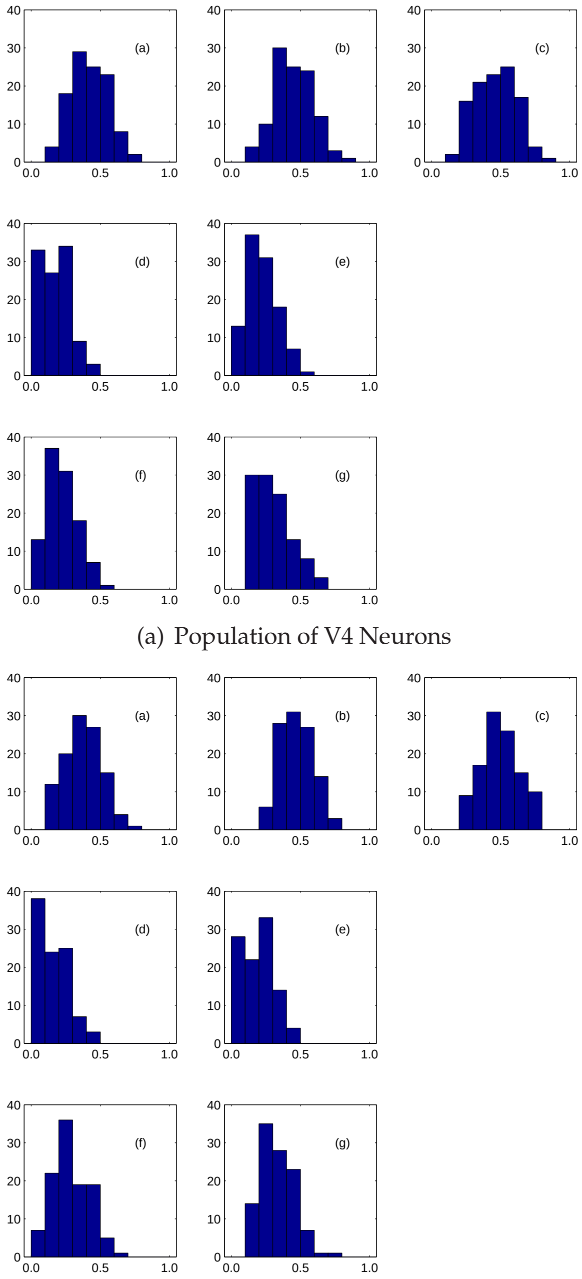
The 2-D boundary conformation domain represents the contour elements of each stimuli in a curvature × angular position space. The 4-D boundary conformation domain contains, in addition to the same curvature × angular position space as the 2-D boundary conformation space, two adjacent curvature dimensions (*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 × elongation tuning functions (f) and 3-D axial orientation × length × 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.

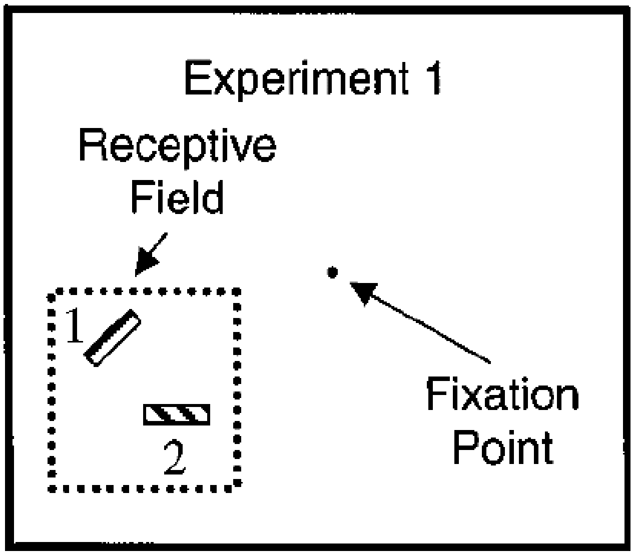


Figure 3-5: The two-bar stimulus experiment by Reynolds *et al.* (modified from [Reynolds et al., 1999]). The monkey here is attending away from the receptive field of the neuron being recorded. Reynolds *et al.* found that the addition of a second stimulus presented within the receptive field of a V4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

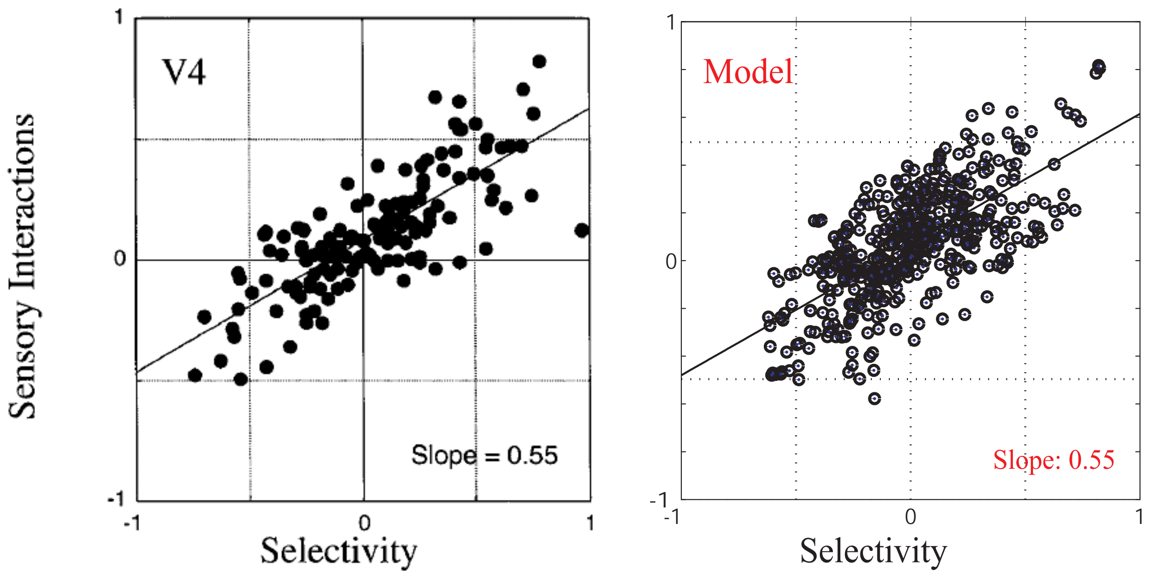


Figure 3-6: The model exhibits a behavior to the two-bar stimuli presentations very similar to the V4 neurons in the absence of attention. The summary of V4 neural responses, adapted from Fig. 5 in [Reynolds et al., 1999], is shown on the left. The addition of a stimulus moves the response toward the response to that stimulus alone, *i.e.*, the response to the clutter condition lies between the responses to the individual stimuli.

1. the reference stimulus appearing in position one alone;
2. the probe stimulus appearing in position two alone;
3. the reference stimulus appearing in position one together with the probe stimulus at position two.

Each unit response was normalized by dividing all responses by the maximal response of the unit across all conditions. As in [Reynolds et al., 1999] we computed several indexes:

- A selectivity index SE_i:

  $$SE_i = PROBE_i - REF,$$

  where $PROBE_i$ is the normalized response of the unit to the reference stimulus and $PROBE_i$ the normalized response of the unit to the i^{th} probe. This was computed for each of the probes thus yielding 16 selectivity values for each unit. This selectivity index can range from -1 to $+1$, with negative values indicating that the reference stimulus elicited the stronger response, a value of 0 indicating identical responses to reference and probe, and positive values indicating that the probe stimulus elicited the stronger response.

- A sensory interaction index SI_i:

  $$SI_i = PAIR_i - REF,$$

  where $PAIR_i$ is the normalized response to the pair composed of the reference stimulus and the i^{th} probe stimulus. The selectivity index also takes on values from -1 to $+1$. Negative values indicate that the response to the pair was smaller than the response to the reference stimulus (*i.e.*, adding the probe stimulus suppressed the neuronal response). A value of 0 indicates that adding the probe stimulus had no effect on the neuron's response. Positive values indicate that adding the probe increased the neuron's response.

C TE and the Model

One of the key feat of the original model [Riesenhuber and Poggio, 1999a] was its ability to duplicate the tuning and invariance properties of the view tuned units from TE/AIT [Logothetis et al., 1995]. To ensure that the model remains consistent with these data, we probed model S_4 units with the paperclip stimuli as in the physiology experiment. As in [Riesenhuber and Poggio, 1999a], we used 80 out of the set of 200 paperclip stimuli (20 targets, 60 distractors) used in [Logothetis et al., 1995].

C.1 Methods

To assess the degree of invariance to stimulus transformations, we used a paradigm similar to the one used in [Logothetis et al., 1995; Riesenhuber and Poggio, 1999a], in which a transformed (rescaled or rotated in depth) target stimulus is considered properly recognized in a certain presentation condition if the S_4 tuned to the original target (default size and view), responds more strongly to its presentation than to the presentation of any distractor stimulus. This measures the hit rate at zero false positives.

To measure the invariance properties of the S_4 units to translation, we trained one S_4 unit for each of the 20 target paperclips presented in the center of the image. During the test period, we probed the response of each S_4 unit to its preferred stimulus in eight possible quadrants (± a random shift within the quadrant, see Fig. 3-7(a)) as well as distractors. To measure the functional receptive field of each S_4 unit, we compared the response of the unit to its preferred stimulus at each location and compared it to the response to a distractor in the center of the visual field.

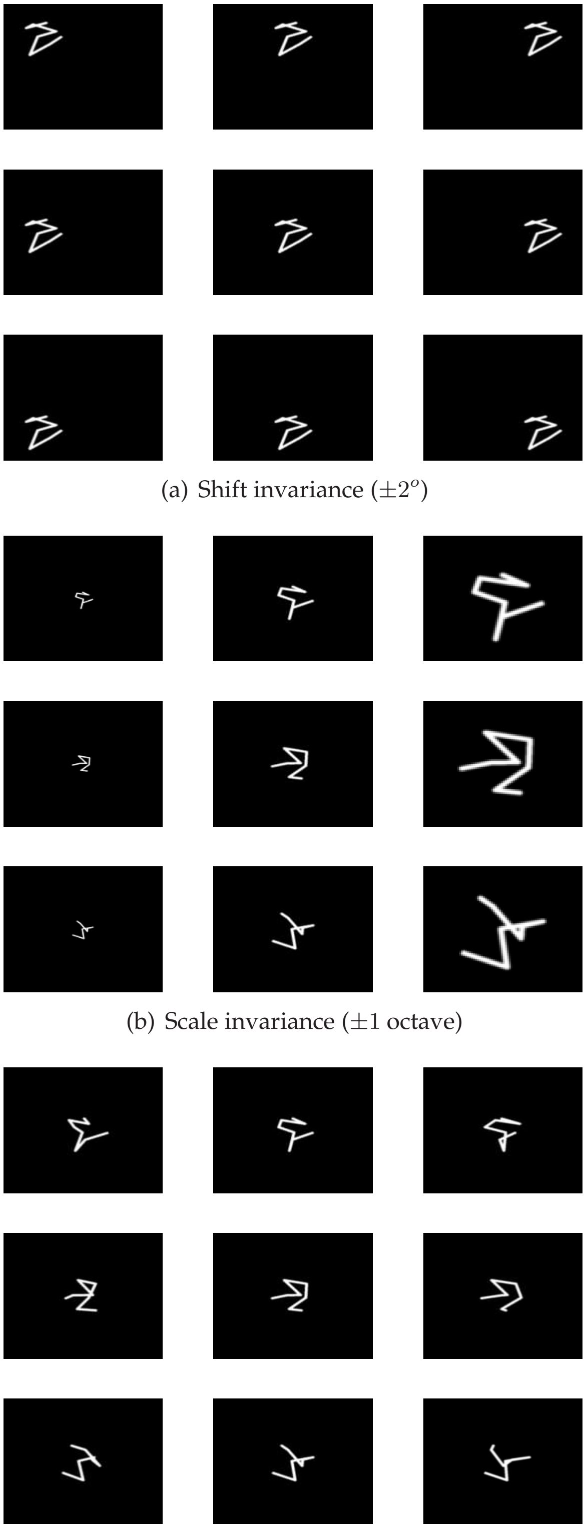
To examine the invariance of the S_4 units to scale, we trained one S_4 unit to each of the 20 target paperclips at the original size (at the center of the visual field). During the test period, we compared the response of each S_4 units to its preferred stimulus at at different sizes (in $1/4$ octave steps, see Fig. 3-7(b)). To examine the invariance of the S_4 units to pose, we trained one S_4 unit to each of the 20 target paperclips at a reference view (denoted 0^o, positioned at the center of the input image). During the test period we probed the response of each S_4 unit with its preferred stimulus at different orientations $\pm 50^o$ from the reference by steps of 4^o, see Fig. 3-7(c) as well as distractors.

C.2 Results

We confirmed that the range of invariance of the S_4 units is well within the range of the view-tuned units in AIT from [Logothetis et al., 1995] (see also [Riesenhuber and Poggio, 1999a]). Model S_4 units exhibited an average position invariance of $\pm 2^o$ and a scale invariance of ± 1 octave. The invariance to 3-D pose was $\pm 20^o$. Also note that in the present version of the model (unlike the original one [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 this benchmark paperclip recognition experiment. Thus the fitting of those paperclip data by the new model is even more remarkable than in the original HMAX case. Details about this experiment can be found in [Serre and Riesenhuber, 2004].

D Discussion

In addition to the comparisons we described in this Chapter, the model has been shown to be qualitatively and quantitatively consistent with several other properties of cells. For instance, the earlier model by [Riesenhuber and Poggio, 1999a] was shown to be compatible with data from PFC [Freedman et al., 2003] as well as several fMRI and psychophysical data [Riesenhuber et al., 2004]. For instance, the model predicts (see [Serre et al., 2005a]), at the C_1 and C_2 levels respectively, the max-like behavior of a subclass of complex cells in V1 [Lampl et al., 2004] and V4 [Gawne, 2000]. [Cadieu, 2005] showed that it is possible to fit individual V4 neurons with model C_2 units (with a very simple greedy fitting approach)



(a) Shift invariance ($\pm 2^o$)

(b) Scale invariance (± 1 octave)

(c) Pose invariance ($\pm 20^o$)

Figure 3-7: Paperclip stimuli used to test the tuning and invariance properties of the model S_4 units as in the monkey physiology experiment [Logothetis et al., 1995].

and to accurately predict their responses across different stimulus sets. For instance, the fitting procedure can be performed on a stimulus set A, *e.g.,* boundary conformations, and still predict the neural responses on another stimulus set B, *e.g.,* 2-spot reverse correlation maps. The model also accounts for the experimental recordings in IT during presentation of multiple objects and read-out from C_{2b} units in the model predicted [see Serre et al., 2005a, section 4.3] recent read-out experiments in IT [Hung et al., 2005], showing very similar selectivity and invariance for the same set of stimuli.

Thus far the model has been successful in making quantitative predictions from V1 through V4, IT and PFC. This strongly suggests that the theory provides an important framework for the investigation of visual cortex.

Notes

[1] Interestingly, sweeping edges, optimal bars and Cartesian gratings gave very similar tuning curves for model units (see Fig. 2-4). This constitutes an important sanity check as different groups tend to use different stimuli to assess the tuning properties of cortical cells.

[2] The correlation coefficients in Fig. 3-4 were found using the same nonlinear fitting procedures with different tuning functions as described in Fig. 9 of [Pasupathy and Connor, 2001]. There are some small numerical differences between our results and those of Pasupathy and Connor. The discrepancies may be due to the minor differences in normalization of the V4 responses (*e.g.,* we linearly scaled the V4 data, courtesy of Pasupathy and Connor, to lie between 0 and 1), differences in conventions for extracting parameters (curvature, edge orientation, axial orientation) from the stimuli, and differences in the nonlinear fitting routines (*e.g.,* number of initial points).

[3] The model assumes that there are *simple* (S_2) and *complex* (C_2) computational units which differ in their translational and scale invariance properties. The available data from V4 suggests that most of the reported results from recording experiments are from C_2-like cells (cells with a range of translation invariance that cannot be attributed to the range of invariance from V1 complex cells). The model predicts the existence of S_2-like cells. They

may at least in part be present in area V2 and feed directly to the C_2-like cells in area V4. We do not think there is enough evidence so far for ruling out the presence of *simple* and *complex* cells in V4 (the difference would be mostly in the larger range of invariance to position and scale for C_2 cells than S_2 cells).

Acknowledgments

The work in Section A and C was done in collaboration with Maximilian Riesenhuber (Georgetown). The work in section B was done in collaboration with Charles Cadieu and Minjoon Kouh at CBCL. We would like to thank Anita Pasupathy and Ed Connor for their help in reconstructing the boundary conformation stimuli, re-implementing the analysis methods, and for sharing their data [Pasupathy and Connor, 2001].

Chapter 4

Performance on Natural Images

In Chapter 3, we showed that model units agree with neural data from V1, V4 and IT. In particular, we showed that a dictionary of shape-components, learned from natural images during a developmental-like learning stage in which model units become tuned to patches of natural images, seem to quantitatively account for the tuning properties of V4 cells on standard stimuli (gratings, boundary conformations and two-bar stimuli). For a theory of object recognition in cortex to be successful, it should also be able to perform robust invariant recognition in the real-world. Here we report on the model performance on several databases of photo-realistic picture images of objects in their natural environment (*e.g.,* in clutter). Images are unsegmented and both the learning and the recognition stages have to cope with clutter. Here we 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.

In Section A, we first evaluate the performance of the model on several categorization tasks with a large database of objects called the *CalTech-101* object database. We also provide experimental simulations that evaluate the robustness of the model to various simplifications in the circuits that approximate the two key operations in the model, *i.e.,* the TUNING and MAX operation (see Chapter 1). In Section B, we compare the performance of the model to several benchmark AI recognition systems. We finally discuss possible implications for biological vision in Section C.

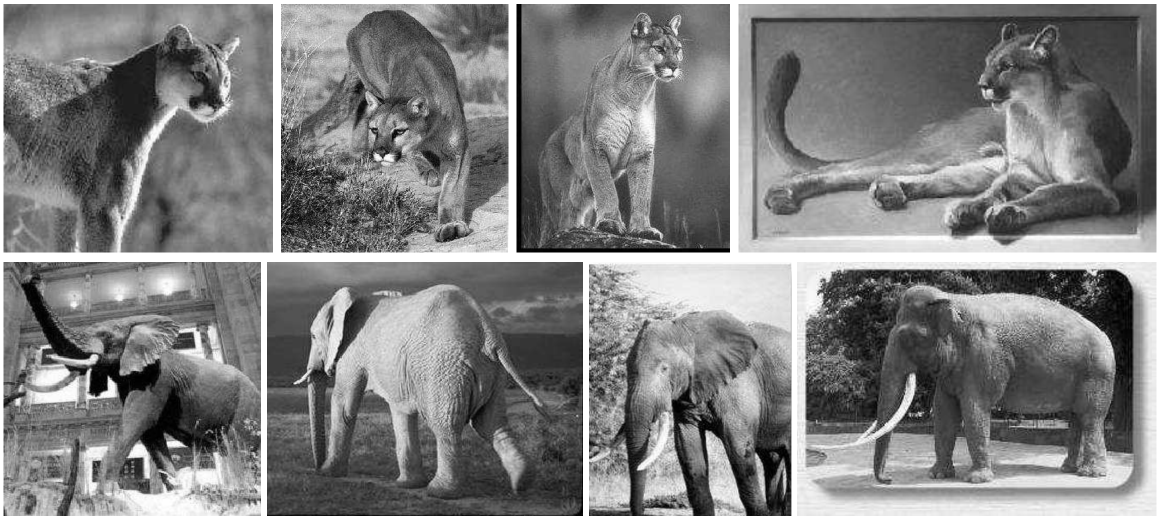


Figure 4-1: Representative images from the *CalTech-101* object database [Fei-Fei et al., 2004] (cougar and elephant categories displayed). The challenge for a vision system is to cope with the drastic changes in the object appearance (*e.g.,* pose, shape, texture, size) as well as changes in clutter and illumination.

A Robustness of the Model

A.1 Performance on the CalTech-101 Database

The CalTech-101 database contains images of objects organized into 101 different categories. Each category contains $\approx 40 - 800$ images with most categories having ≈ 50 images. The size of each image is roughly 300×200 pixels but to speed-up processing time, we rescaled all images to be about half the original size (more precisely images were rescaled to be 140 pixels in height). Images were collected from the web by Fei-Fei and colleagues (see [Fei-Fei et al., 2004]) using a search engine[1]. The database constitutes a challenge for a vision system as it contains images from many different object categories with large variations in shape, clutter, pose, illumination, size, *etc* . Some of the objects are highly "deformable" (*e.g.,* animals appearing in any pose). Representative images from two animal categories (elephant and cougar) are shown in Fig. 4-1.

Importantly, although images in the database have been recently annotated to provide the outline of objects, we did not use these annotations. That is, the set of images used to train and test the model was unsegmented (*i.e.,* objects were embedded in clutter). Also, while images in the database contain color information, as a pre-processing step, we converted all images to gray-scale. The database is becoming increasingly popular, which makes it very useful for performing standardized comparisons between different approaches [Fei-Fei et al., 2004; Berg et al., 2005; Serre et al., 2005b; Holub et al., 2005a,b; Grauman and Darrell, 2005]. Altogether the database constitutes an interesting challenge for a neurobiological model of object recognition.

Methods: The core part of the model, *i.e.,* the dictionary of shape-components corresponding to units from V4 to TEO used in this experiment was obtained with the procedure described in Chapter 2 (Section B). This developmental-like learning stage sets the preferred stimulus of the S units (*i.e.,* their synaptic weight vector $\mathbf{w}$, see Eq. 1.2) in several layers of the model. As a result, units become tuned to *key image-features* that occur with high probability in natural images. During this unsupervised learning stage, the model is exposed to a few hundred random natural images, unrelated to any categorization task.

The resulting dictionary of features is *generic* in the sense that, as we show below, it can support the recognition of a large variety of object categories. As discussed earlier in Chapter 2, this dictionary of shape-components is redundant and overcomplete and contains features of various complexities. For instance, the simplest features (*i.e.,* a simple combination of V1-like oriented subunits with small range of invariance and corresponding to cells in V4 (see Chapter 3) are computed at the S_2 level, whereas more complex features (*e.g.,* object-part detectors with a larger range of invariance which may be similar to some of the features found in TEO columns [Fujita et al., 1992; Wang et al., 1996; Tanaka, 1996, 1997; Wang et al., 1998; Tanaka, 2003]) are computed higher up in the hierarchy (layers C_3 and C_{2b}).

All the experiments presented here were, however, performed with an earlier implementation of the model which is simpler than the one described in Chapter 2. The precise architecture is illustrated in Fig. 4-2. Some of the routes are missing (see translucent components of the model in Fig. 4-2). In particular, the route from $S_2 \rightarrow C_2 \rightarrow S_3 \rightarrow C_3$ is absent. As we confirmed experimentally, while the full architecture performs significantly better than this simpler implementation (*e.g.,* in Chapter 5, the complete dictionary of shape-components is necessary to account for the level of performance of human observers), we believe that results would remain *qualitatively* similar with the full architecture. Additionally, this “simplified” implementation presents the advantage that it runs significantly faster than the full implementation and therefore allows more experiments to be completed.

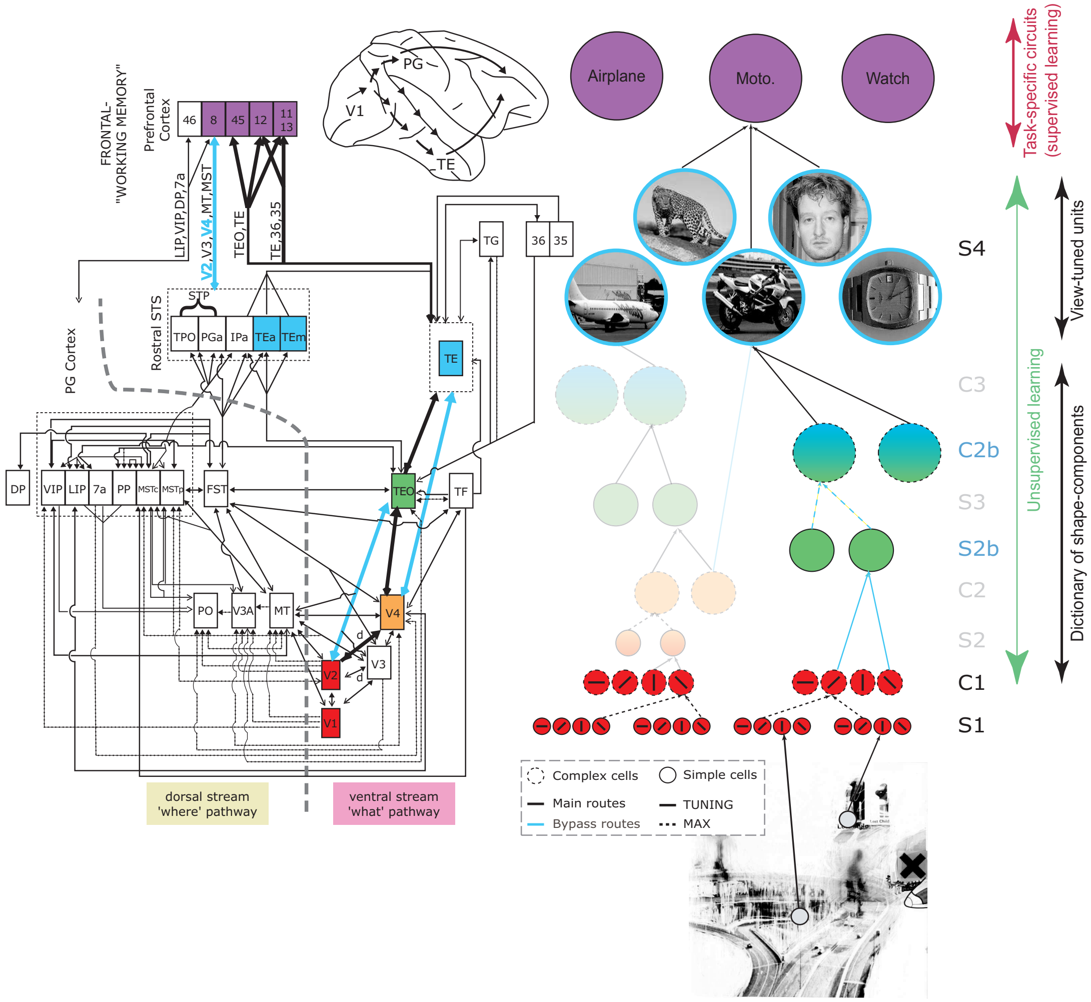


Figure 4-2: Schematic of the "simplified" architecture used here (based on an earlier model implementation). Some of the routes are missing (indicated in translucent). During training, examples are stored at the level of the S_4 units which provide a *holistic* and *view-based* representation of target objects. At the top of the hierarchy, PFC *classification* units combine the response of several S_4 units. The PFC *classification* units perform simple binary classification tasks (object present / absent). During training (which is the only supervised learning stage in the model), one PFC *classification* unit is learned for each of the object categories). By comparing the response of a single PFC *classification* unit in the presence and absence of its associated target object, the model can be evaluated on a detection task. By considering the response of all the *classification* units and assigning to the input image the label of the *classification* unit which is maximally activated, the model can be evaluated on a more challenging N-alternative forced (see Fig. 4-4).

To train the model to perform different categorization tasks (*e.g.*, face present / absent), we trained the *task-specific* circuits at the top of the hierarchy (*i.e.*, the S_4 and PFC units). This was done by performing random splits (between images used for training and images used for testing the model) over each image dataset. That is, for each image category, we selected a variable number N_{Tr} of images for training and up to $N_{Te} = 50$ images for testing (from the remaining images, not used for training). This procedure is also called *leave-out* procedure (see [Devroye et al., 1996]) and has been shown to provide a good estimate of the expected error of a classifier.

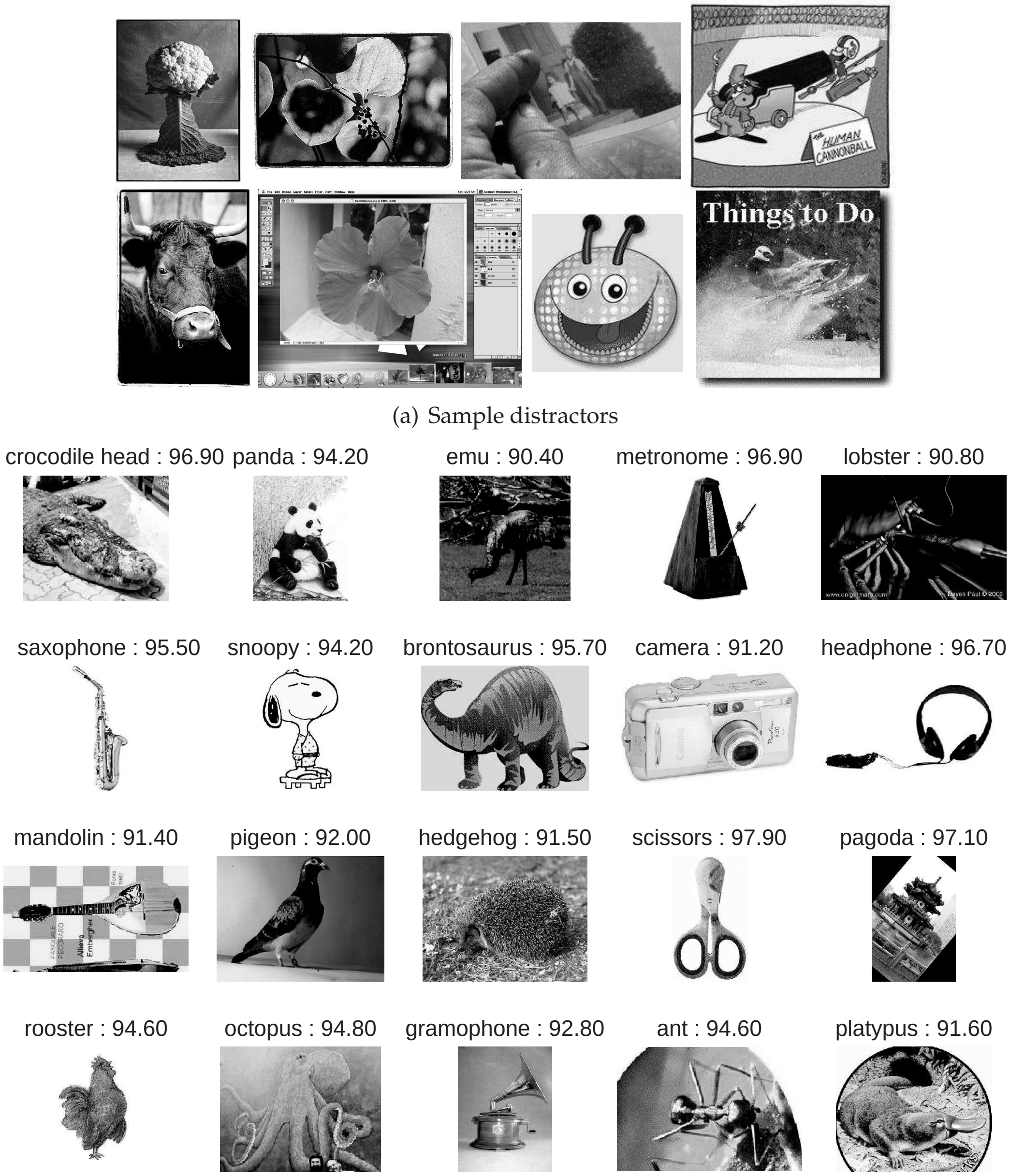
As described in Chapter 2, the task-specific circuits of the model are trained in a supervised way and in two steps. For each object category:

1. Typical examples (*i.e.,* $\approx 25\%$ of the training set) from the target object set are stored at the level of the S_4 units (one S_4 per example to be stored). The S_4 units provide a *holistic* and *view-specific* representation of familiar objects [Logothetis et al., 1995] at the level of AIT (see Chapter 1, Section B).

2. One PFC *classification* unit in PFC is trained, *i.e.,* its synaptic weights $\mathbf{c}$ (see Eq. 2.1) are adjusted so as to minimize the classification error on the training set (Eq. 2.2).

During the test period, we compute the error of the PFC classification units on the set of test images (not used for training). For each image category we plot an ROC curve [2] and estimate the area under the curve as the performance measure for the model (as in [Fei-Fei et al., 2004]). The procedure is re-iterated 10 times for each category: Each time we generate a different training and test set at random, train one PFC classification unit with the procedure described above, then evaluate the performance on the test set and evaluate the area under the ROC curve. We report the average performance across these 10 random runs.

Sample results: Fig. 4-3(b) shows some typical results from sample object categories. The performance of the model is remarkable given the fairly small number of training examples used (< 100). Typical computer vision systems generally use thousands of training examples [Sung, 1996; Osuna, 1998; Schneiderman and Kanade, 2000; Heisele et al., 2002]. Indeed, in [Serre et al., 2005b], we found that the size of the training set could be further reduced and that reasonably good performance could be obtained with very few training examples (just $3-6$ positive training examples). The reader may refer to [Serre et al., 2005b, 2006b] and Appendix B for more extensive simulations and results on the *CalTech-101* object dataset.



(a) Sample distractors

(b) Sample results on the *CalTech-101* object database

Figure 4-3: a) Typical distractors used to evaluate the performance of the model on an object present *vs.* absent task. **b**) Sample results obtained with the model on the *CalTech-101* object database. The paradigm used here to evaluate the performance of the model is standard (see [Fei-Fei et al., 2004] for instance). Like observers in rapid-categorization tasks, the model classifies a particular stimulus as object (*e.g.,* an animal) present or absent. For all categories, the set of distractors was sampled at random from a separate *background* image set **a)** containing a large number of scenes (at different scales) that do not contain any of the target objects (see [Fei-Fei et al., 2004]). Each thumbnail illustrates a typical example from the image set and the number above corresponds to the average performance across 10 *random runs*. In each run, the model is trained using a small number of labeled examples (see text) and tested on a separate set.

A.2 Approximating the Key Computations

In the experiment described above, the model implementation used relies on exact computations of the two key operations, *i.e.,* an exact Gaussian TUNING (Eq. 1.2) and an exact MAX operation (Eq. 1.1). Yet, this is unrealistic and as discussed in Chapter 1, biophysically plausible circuits would, at best, implement crude approximations of the two operations. Another idealization from the previous experiment comes from the use of *continuous analog* unit responses. As illustrated in Fig. 1-5, circuits implementing the key operations are likely to rely on quantized values provided by *computational modules*, *i.e.,* groups of n equivalent cells (see Chapter 1). The level of quantization, in turn, is determined by the number n of equivalent units within each module. The precise nature of such modules and the number of cells they contain is yet to be determined both experimentally and theoretically.

While it will be important in the future to test the model with more realistic biophysical circuits of the key operations (the advent of large-scale neural architectures such as *Blue Brain* will certainly provide the computational machinery necessary), we start here more modestly and test the robustness of the model to "simplifications" which take the form of various approximations in the circuits involved in the TUNING operation.

First we test how critical is the use of analog continuous responses at the level of individual units. While we think that several levels of quantizations are certainly necessary in the lowest levels (at the S_1 and C_1 levels), we suggested in Chapter 1, based on the anatomy and physiology of the ventral stream, that the number n of units in individual computational module may decrease along the hierarchy, with units in the higher-most layers behaving essentially as switches (being either ON or OFF). We here test this hypothesis more directly by binarizing the response of the units in an intermediate layer of the model, the C_{2b} layer (corresponding to cortical area TEO) which gives inputs to the S_4 units (corresponding to the view-tuned units in TE).

Here, we also test the robustness of the model to different TUNING approximations at the S_4 level and compare the performance of a) an exact Gaussian function (Eq. 1.2), b) a normalized dot-product (Eq. 1.3) and c) a simple dot-product (*i.e.,* when dropping the normalization term in Eq. 1.3).

Methods: All experiments in the following were performed on a subset of the *CalTech-101* object dataset. We first selected image categories that contained at least 150 examples so as to perform several random runs with 100 training and 50 test examples selected at random for every run. This lead to five categories, *i.e.,* faces, leopards, motorcycles, airplanes and watches as well as an additional set of distractors from the *background* category.

To test the model dependency on the use of analog *vs.* binary values we compared the performance of the standard model implementation (*i.e.,* analog values) with a model implementation that relies on binary unit responses in one of the intermediate (C_{2b}) layers, which then provides inputs to the S_4 units. We first calculated a response threshold θ for each of the C_{2b} units such that the corresponding unit would be active on $P\%$ of the entire training set and inactive on the remaining $100 - P\%$. We experimented with different values of P, *i.e.,* $P = 10\%,\ 30\%,\ 60\%$. The performance of the model was evaluated with two different paradigms:

- A target present/absent paradigm (chance level 50%) for each of the five classes separately, *i.e.,* each recognition task was evaluated as an independent binary classification problem. Distractors were randomly sampled from the same separate set of distractors ("background" category);

- A N-alternative forced choice paradigm (where $N = 5$ is the total number of classes, chance level 20%), *i.e.,* each image presented has to be classified by the model as either one of five possible categories.

Fig. 4-4 and 4-5 show simulation results for $P = 30\%$. When the number of afferents is large enough (> 100), the loss in performance induced by the binarization of the units becomes negligible. Note that we found qualitatively similar results for $P = 10\%$ and $P = 60\%$ (not shown) and observed a large drop in performance for higher values of P.

These results show that the model does not rely critically on exact computations and can rely on approximations. This, in turn, suggests that the performance obtained with *idealized* operations may generalize to approximate operations performed in biophysically plausible circuits. The results of the simulations give support to the hypothesis from Chapter 1. That is, the size of the computational modules (*i.e.,* the number n of equivalent units that receive the same inputs and encode the same feature dimension) may decrease along the hierarchy with units in the top-layers behaving essentially like switches (being either ON or OFF).

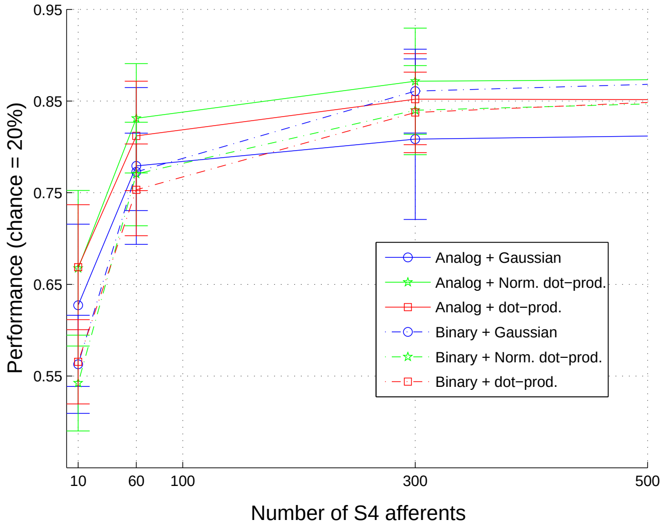


Figure 4-4: Robustness of the model to approximate computations at the S_4 level. The model (see Fig. 4-2) is evaluated on a challenging 5-alternative forced choice (chance level 20%). The model maintains a high recognition rate even when unit responses at the C_{2b} level are binary and when the TUNING operation is approximated, *i.e.*, exact Gaussian (see Eq. 1.2) *vs.* normalized dot-product (Eq. 1.3) *vs.* simple dot-product (*i.e.*, the normalization term is dropped), see Chapter 1 for details.

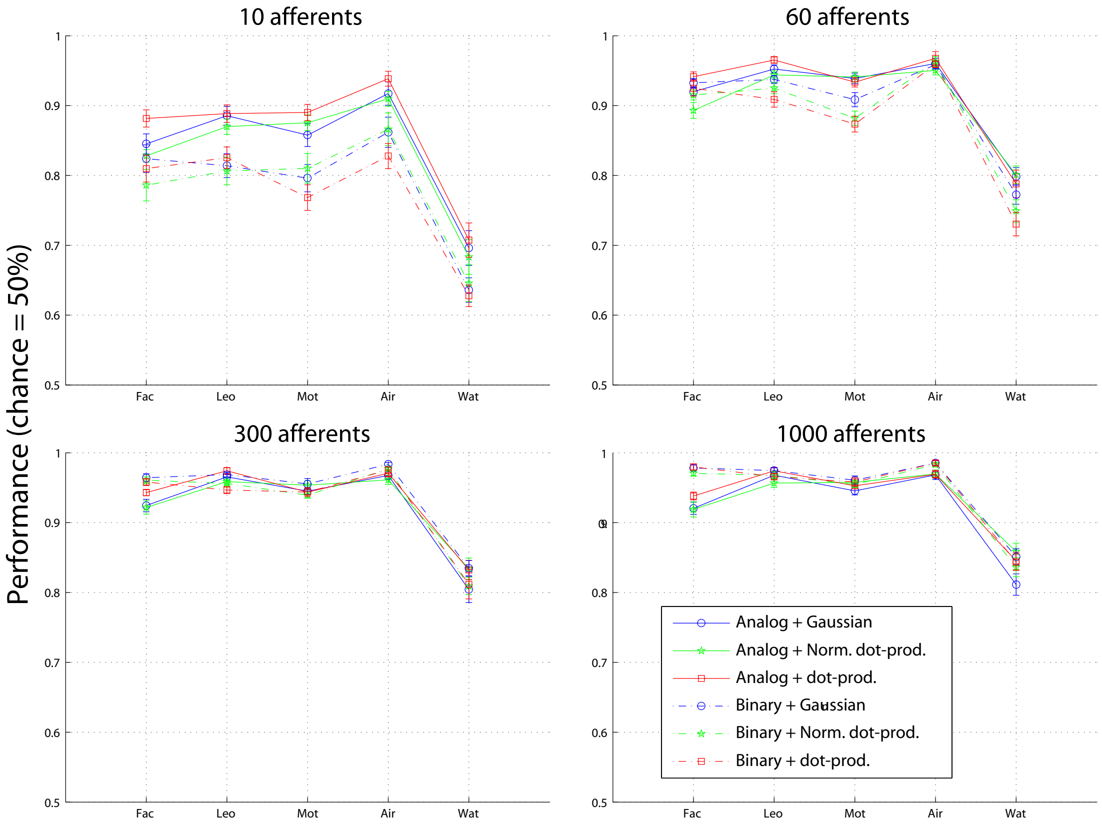
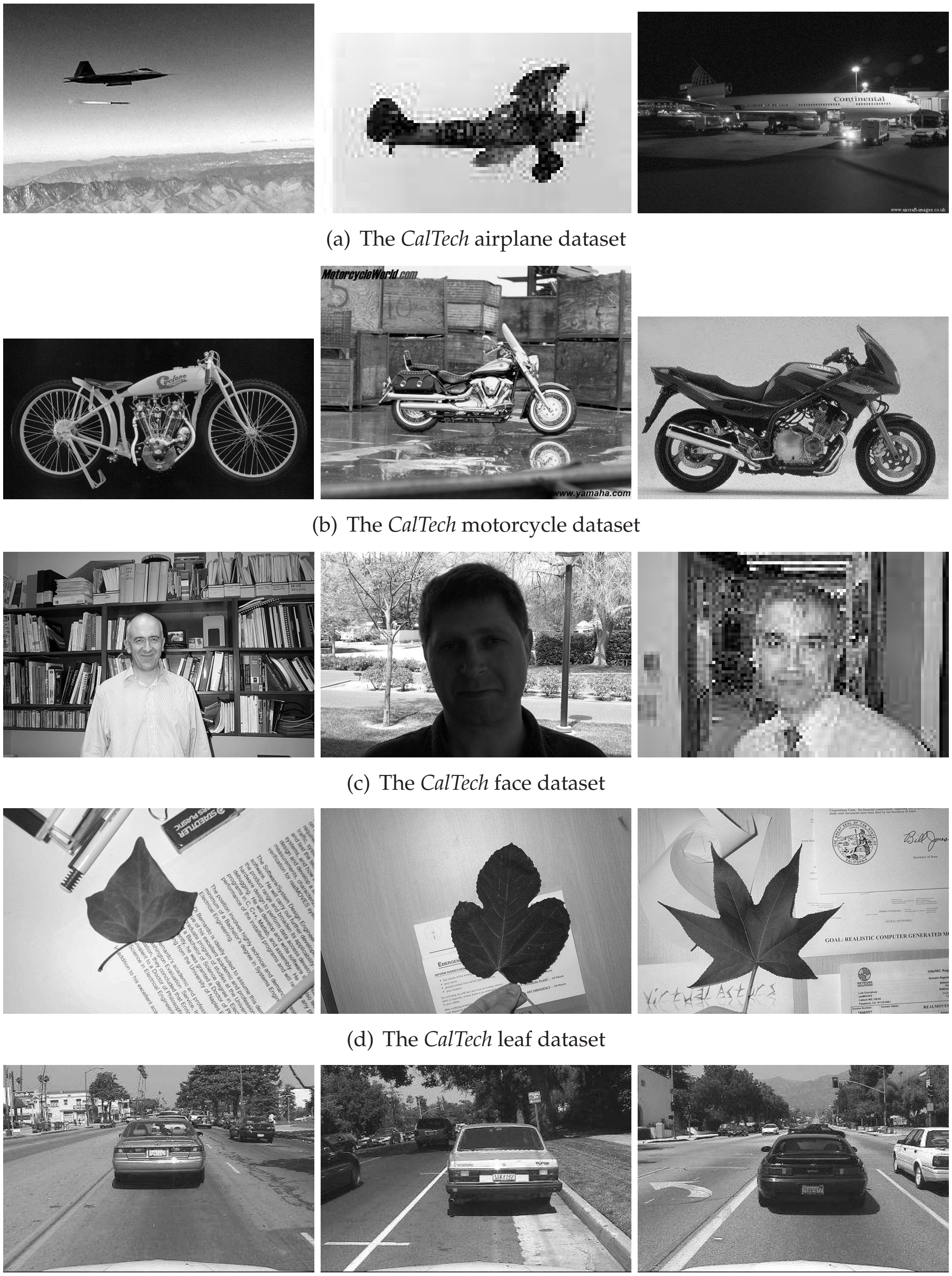


Figure 4-5: Robustness of the model to approximate computations at the S_4 level. The model is evaluated on a simpler object present/absent recognition task (for each class independently, *i.e.*, face, leopard, motorcycle, airplane and watch *vs.* background images). Chance level is 50%. As for the multi-class classification problem (see Fig. 4-4), the model maintains a high recognition rate despite various simplifications in the computations performed.



(a) The *CalTech* airplane dataset

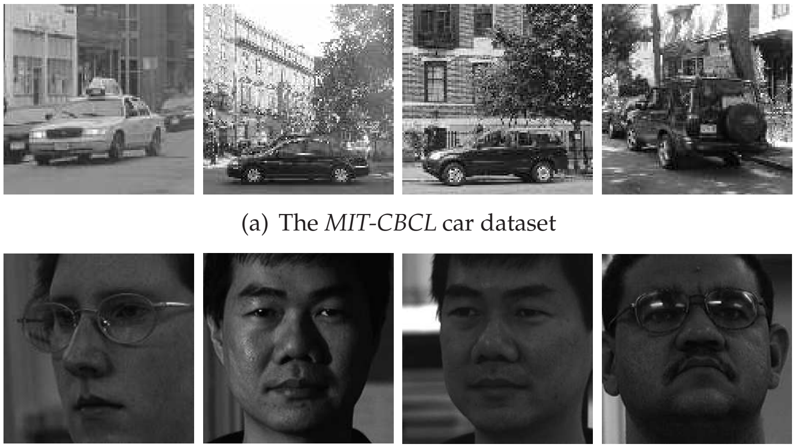
(b) The *CalTech* motorcycle dataset

(c) The *CalTech* face dataset

(d) The *CalTech* leaf dataset

(e) The *CalTech* car dataset

Figure 4-6: The *CalTech* datasets used to compare the model to other benchmark AI systems [Weber et al., 2000b; Fergus et al., 2003].



(a) The *MIT-CBCL* car dataset

(b) The *MIT-CBCL* face dataset

Figure 4-7: The *MIT-CBCL* datasets used to compare the model to other benchmark AI systems [Heisele et al., 2002; Leung, 2004].

B Comparison with Standard AI Recognition Systems

For a more rigorous and objective evaluation, in this section, we compare the performance of the model to other AI recognition systems. For this comparison we used standard datasets (see Fig. 4-6 and Fig. 4-7) from two vision groups, *i.e.,* two *MIT-CBCL* datasets from our own group and five (*CalTech-5*) datasets from the CalTech vision group.

CalTech-5: We consider five databases from the CalTech vision group[3], *i.e.,* frontal-face, motorcycle, rear-car and airplane datasets from [Fergus et al., 2003], as well as the leaf dataset from [Weber et al., 2000b] (see Fig. 4-6 for examples). On these datasets, we used the same fixed splits as in the corresponding studies whenever applicable and otherwise generated random splits. All images were rescaled to be 140 pixels in height (width was rescaled accordingly so that the image aspect ratio was preserved) and converted to grayscale.

MIT-CBCL: This includes a near-frontal ($\pm 30^\circ$) face dataset [Heisele et al., 2002] and a multi-view car dataset from [Leung, 2004] (see Fig. 4-7). The face dataset contains about 6,900 positive and 13,700 negative images for training and 427 positive and 5,000 negative images for testing. The car dataset contains 4,000 positive and 1,600 negative training examples and 1,700 test examples (both positive and negative). Although the *benchmark* algorithms were trained on the full sets and the results reported accordingly, our system only used a subset of the training sets (500 examples of each class only).

These two MIT-CBCL datasets are challenging: The face patterns used for testing are a subset of the CMU PIE database [Sim et al., 2001] which contains a large variety of faces under extreme illumination conditions (see [Heisele et al., 2002]). The test non-face patterns were selected by a low-resolution LDA classifier as the most similar to faces (the LDA classifier was trained on an independent 19×19 low-resolution training set). The car database includes a wide variety of vehicles, including SUVs, trucks, buses, *etc* , under wide pose and lighting variations. Random image patterns at various scales that were not labeled as vehicles were extracted and used as a negative test set.

Methods: For this comparison, we also used the earlier (simpler) implementation of the model (see Fig. 4-2) which corresponds to the route projecting from $S_1 \rightarrow C_1 \rightarrow S_{2b} \rightarrow C_{2b}$. Again, the performance of the full architecture which include a richer dictionary of shape-components, tends to be significantly higher than the performance of this simpler (incomplete) implementation. Therefore the results reported here are likely to constitute only a lower bound on the system performance.

Also, for a fair comparison with the benchmarks and in order to emphasize the contribution of the feature representations rather than the classification modules, we passed the response of the C_{2b} units directly to a linear classifier. This allows for a more rigorous comparison at the representation-level (model C_{2b} units *vs.* computer vision features such as SIFT [Lowe, 1999], component-experts [Heisele et al., 2002; Fergus et al., 2003; Fei-Fei et al., 2004], or fragments [Ullman et al., 2002; Torralba et al., 2004]).

Results: Table 4.1 summarizes our main results. The model performs surprisingly well, better than all the systems we have compared it to thus far. In Appendix B we provide additional results and comparisons to other types of features (*e.g.,* SIFT features [Lowe, 2004]). Altogether the results suggest that the model can outperform other AI systems in different conditions such as, recognition of objects in clutter, recognition of objects in segmented scenes (in combination with a scanning approach, see [Serre et al., 2006b]) and for the recognition of shape-based (*e.g.,* car, face, *etc*) as well as texture-based (*e.g.,* tree, building, *etc*) objects. Details about these comparisons may be found in [Serre et al., 2004b, 2005b; Bileschi and Wolf, 2005; Serre et al., 2006b].

Datasets			AI systems	Model
(CalTech)	Leaves	[Weber et al., 2000b]	84.0	97.0
(CalTech)	Cars	[Fergus et al., 2003]	84.8	99.7
(CalTech)	Faces	[Fergus et al., 2003]	96.4	98.2
(CalTech)	Airplanes	[Fergus et al., 2003]	94.0	96.7
(CalTech)	Motorcycles	[Fergus et al., 2003]	95.0	98.0
(MIT-CBCL)	Faces	[Heisele et al., 2002]	90.4	95.9
(MIT-CBCL)	Cars	[Leung, 2004]	75.4	95.1

Table 4.1: The model *vs.* other AI benchmark recognition systems. For the *CalTech-5* datasets (*i.e.,* leaf, car, face, airplane, and motorbike), the objects are presented in clutter and all the systems are trained and tested on unsegmented images. The benchmark systems are the *constellation* models by Perona and colleagues [Weber et al., 2000b; Fergus et al., 2003], which rely on part-based generative models of the object. For the *MIT-CBCL* face dataset we compare with a hierarchical SVM-based architecture that was, by itself, shown to outperform several other face-detection systems [Heisele et al., 2001c, 2002]. For the *MIT-CBCL* car dataset we compared to a system by [Leung, 2004] that uses fragments [Ullman et al., 2002] and similar to [Torralba et al., 2004]. The performance measure reported is the *performance at equilibrium* which corresponds to the error rate for which the miss rate is equal to the false-alarm rate, see [Serre et al., 2005b, 2006b] for details.

C Discussion

To summarize, we described experiments which showed that:

- An implementation of the theory described in Chapter 2 is able to handle the invariant recognition of many different object categories with the same basic dictionary of shape-components.

- The model performs very well on simple detection tasks (*i.e.,* object present / absent) as well as more challenging N-alternative forced choice recognition tasks.

- The model does not seem to depend critically on the exactitude of the key computations (at least in the top layers) and various approximations can still support robust invariant recognition. In particular, because along the hierarchy units receive more and more inputs, the TUNING operation may not need to be exact and indeed a simple dot-product (which approximate well Gaussian tuning in high dimensional space) may well be sufficient. Further work will be needed to test the robustness to approximations to the key computations in lower stages.

  We also found that in the top stages of the model, a graded responses along particular feature dimensions may not be needed (as we originally anticipated) and that binary unit responses (*i.e.,* simply *on* or *off*) may be sufficient to support robust invariant recognition.

We found that the level of performance achieved by the model is far from trivial and that indeed the model can outperform other AI systems (this is the case on all the tests we have performed thus far). Additionally, recent work is already suggesting that the performance of the model can be further improved. On the *CalTech-101* database, using a (non-biological) multi-class SVM on all 101 categories with 15 training examples per class averaged over 10 repetitions, we obtained $44\% \pm 1.14\%$ correct classification rate [Serre et al., 2006b].

By enlarging the dictionary of shape-components and computing additional *gestalt*-like features (*e.g.,* good-continuity detectors, circularity detectors and symmetry detectors) within the same framework, Wolf & Bileschi obtained $\approx 51.2\% \pm 1.2\%$ correct [Wolf et al., 2006; Bileschi and Wolf, 2006]. Mutch & Lowe reported 56% correct by applying a non-biological feature selection method [Mutch and Lowe, 2006]. Some of the best (non-biological) systems include the system by [Holub et al., 2005b] ($\approx 44\%$ correct) and the system by [Berg et al., 2005] (45% correct). To date results obtained within the framework of the theory constitute the state-of-the-art.

Typically, previous models of object recognition have been tested on idealized stimulus set (of the type used in physiology labs) such as simple combinations of bars, or faces presented on a blank background. For instance, using the same paperclip stimuli as used in a psychophysics [Logothetis et al., 1994] and physiology experiment [Logothetis et al., 1995], Riesenhuber & Poggio showed that an earlier implementation of the model presented here, was able to account quantitatively for the tuning properties of the view-tuned units in inferotemporal cortex, which respond to images of the learned object more strongly than to distractor objects, despite significant changes in position and size [Riesenhuber and Poggio, 1999a].

The capacity of the architecture to handle the recognition of a variety of real-world object recognition tasks (*i.e.*, presence of clutter and changes in appearance, illumination, *etc*) provides another compelling plausibility proof for this class of models. This may be the first time that a neurobiological model, faithful to the anatomy and physiology of visual cortex, competes with engineered computer-vision systems.

Indeed, while a long-time goal for computer vision has been to build a system that achieves human-level recognition performance, the state-of-the-art algorithms have been diverging from biology: for instance, some of the best existing systems use geometrical information about the constitutive parts of objects (constellation approaches rely on both appearance-based and shape-based models [Weber et al., 2000b; Fergus et al., 2003; Fei-Fei et al., 2004] and component-based systems use the relative position of the detected components along with their associated detection values [Heisele et al., 2002]). Biology is however unlikely to be able to use geometrical information – at least in the cortical stream dedicated to shape processing and object recognition. The model respects the properties of cortical processing (including the absence of geometrical information) while showing performance at least comparable to the best computer vision systems.

The fact that this biologically-motivated model outperforms more complex computer vision systems might at first appear puzzling. The architecture performs only two kinds of computations (TUNING, equivalent to template matching in computer vision and MAX pooling, also used in computer vision to suppress multiple detections within a neighborhood). Some of the other systems we have compared it to involve complex computations like the estimation of probability distributions [Weber et al., 2000b; Fergus et al., 2003; Fei-Fei et al., 2004] or the selection of facial-components for use by an SVM [Heisele et al., 2002].

It is likely that part of the strength of the model comes from its built-in gradual invariance to position and scale that closely mimics visual cortical processing, which has been finely tuned by evolution over thousands of years. It is also very likely that such hierarchical architecture ease the recognition problem by decomposing the task into several simpler ones at each layer.

Finally it is worth pointing out that the set of shape-component features that is passed to the final classifier is very redundant, probably more redundant than for other approaches. While we showed that a relatively small number of features (about 50) is sufficient to achieve good error rates [Serre et al., 2005b], we have found that the level of performance of the model can be significantly increased by adding many more features. Interestingly, the number of features needed to reach the ceiling ($\approx 1,000 - 5,000$ features, *i.e.,* about the same number of feature columns found by Tanaka and colleagues [Tanaka, 1996], see Chapter 2) is much larger than the number used by current AI systems ($\approx 10 - 100$ for [Ullman et al., 2002; Heisele et al., 2002; Torralba et al., 2004] and $\approx 4 - 8$ for constellation approaches [Weber et al., 2000b; Fergus et al., 2003; Fei-Fei et al., 2004]).

Notes

[1] The *CalTech-101* database is available at:
`http://vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html`.

[2] Using ROC curves to evaluate the performance of a system is common practice in computer vision. By adding a bias term to the output of a classifier (here the PFC units), one can arbitrarily increase or decrease the propensity of the system to classify an image as target or distractor. An ROC curve can be obtained by computing the hit rate and the false-alarm rates of the system for all possible values of the bias. Typically the system performance will range from 0% hit and false-alarm rates for small values of the bias term (*i.e.,* all images classified as distractors) to 100% hit and false-alarm rates for large values of the bias term (*i.e.,* all images classified as targets). Typically meaningful operating ranges are obtained for intermediate values of the bias term. The overall performance of the system can be summarized by the area under the curve.

[3] The *CalTech-5* databases are publicly available at:
`http://www.robots.ox.ac.uk/~vgg/data3.html`.

[4] The source code for the model implementation described in Fig. 4-2 is available at:
`http://cbcl.mit.edu/software-datasets`.

[5] The simpler model implementation used here may correspond to a *bypass* route in cortex (effectively skipping two stages – V4 and TE – in cortex) that projects from V2 directly onto TEO and then from TEO directly onto PFC. In fact, our results suggest that a similar bypass route in cortex could account for some of the fastest reaction times ($120\ ms$) observed in human observers during a forced-choice saccade task [Kirchner and Thorpe, 2005] and which seem irreconcilable with a full processing by the entirety of the ventral stream.

[6] The model, in its present form, does not make use of any special mechanisms for *binding* features together (*e.g.,* synchrony, *etc*) and is therefore with general shape-processing during pre-attentive vision.

Acknowledgments

Part of this Chapter appeared in [Serre et al., 2005b, 2006b]. The evaluation of the model on real-world image datasets was done in collaboration with Lior Wolf and Stan Bileschi at CBCL. I would also like to thank Max Riesenhuber, Jennifer Louie, Rodrigo Sigala and Robert Liu for their contributions in earlier phases of this work.

Chapter 5

Predicting Human Performance in a Rapid Categorization Task

We showed earlier in Chapter 4 that the model is capable of recognizing well complex images and, when tested on real-world natural images, it competes with and sometime even outperforms state-of-the-art computer vision systems on several categorization tasks. It is therefore natural to ask, in this Chapter, whether the model may be able to duplicate human-level performance in complex recognition tasks. This Chapter corresponds to a manuscript in preparation [Serre et al., 2006a].

Abstract

Primates are remarkably good at recognizing objects. The level of performance of the primate visual system and its robustness to image degradations have remained unchallenged by the best computer vision systems despite decades of engineering effort. In particular, the high accuracy of primates in ultra-rapid object categorization [Thorpe et al., 1996] and rapid serial visual processing [Potter, 1975] is remarkable. Given the number of processing stages involved and typical neural latencies, such rapid visual processing is likely to be mostly feedforward [Thorpe et al., 1996; VanRullen and Koch, 2003; Bacon-Mace et al., 2005; Kirchner and Thorpe, 2005].

Yet, so far, no biologically plausible feedforward model of visual cortex has been shown to be capable to perform at human level. Here we show that a specific implementation [Riesenhuber and Poggio, 1999a; Serre et al., 2005a] of a class of feedforward theories of ob-

ject recognition [Hubel and Wiesel, 1968; Fukushima, 1980; Perrett and Oram, 1993; Wallis and Rolls, 1997; Mel, 1997; Hochstein and Ahissar, 2002; Ullman et al., 2002; Thorpe, 2002; Amit and Mascaro, 2003; Wersing and Koerner, 2003] – that extend the Hubel & Wiesel simple-to-complex cell hierarchy and account for many anatomical and physiological constraints – can predict the level and the pattern of performance achieved by humans on a rapid animal *vs.* non-animal categorization task.

A Introduction

Object recognition in cortex is mediated by the ventral visual pathway running from primary visual cortex (V1) through extrastriate visual areas V2 and V4 to inferotemporal cortex (IT, comprising PIT and AIT) and then to prefrontal cortex (PFC) which is involved in linking perception to memory and action. Over the last decade, a number of physiological studies in non-human primates have established several basic facts about the cortical mechanisms of recognition. The accumulated evidence points to several key features of the ventral pathway. From V1 to IT, there is an increase in invariance to position and scale [Hubel and Wiesel, 1968; Perrett and Oram, 1993; Logothetis et al., 1995; Tanaka, 1996; Riesenhuber and Poggio, 1999a] and, in parallel, an increase in the size of the receptive fields [Perrett and Oram, 1993; Tanaka, 1996] as well as in the complexity of the optimal stimuli for the neurons [Desimone, 1991; Perrett and Oram, 1993; Kobatake and Tanaka, 1994]. Finally plasticity and learning are probably present at all stages, and certainly at the level of IT [Logothetis et al., 1995] and PFC. However an important aspect of the visual architecture – the role of the anatomical back-projections abundantly present between almost all of the areas in visual cortex – remains a matter of debate.

It is well known that recognition is possible for scenes viewed in rapid visual presentation that do not allow sufficient time for eye movements [Potter, 1975; Thorpe et al., 1996; Thorpe and Fabre-Thorpe, 2001; VanRullen and Koch, 2003; Bacon-Mace et al., 2005] and in the near-absence of attention [Li et al., 2002]. The hypothesis that the basic processing of information is feedforward is supported most directly by the short times required for a selective response to appear in IT cells [Perrett et al., 1992]. Very recent data [Hung et al., 2005] convincingly show that the activity of small neuronal populations in monkey IT, over very short time intervals (as small as $12.5\ ms$) and only about $100\ ms$ after stimulus onset, contains surprisingly accurate and robust information supporting a variety of recognition tasks. Furthermore, EEG studies [Thorpe et al., 1996] have provided evidence that the human visual system is able to solve an object detection task – determining whether a natural scene contains an animal or not – within $150\ ms$ (see [Kirchner and Thorpe, 2005] for time estimates based on eye movements with a forced-choice saccade task). While this does not rule out the use of local feedback loops within an area, it does suggest that a core hierarchical feedforward architecture may be a reasonable starting point for a theory of visual cortex, aiming to explain the initial phase of recognition which may be called *immediate recognition*.

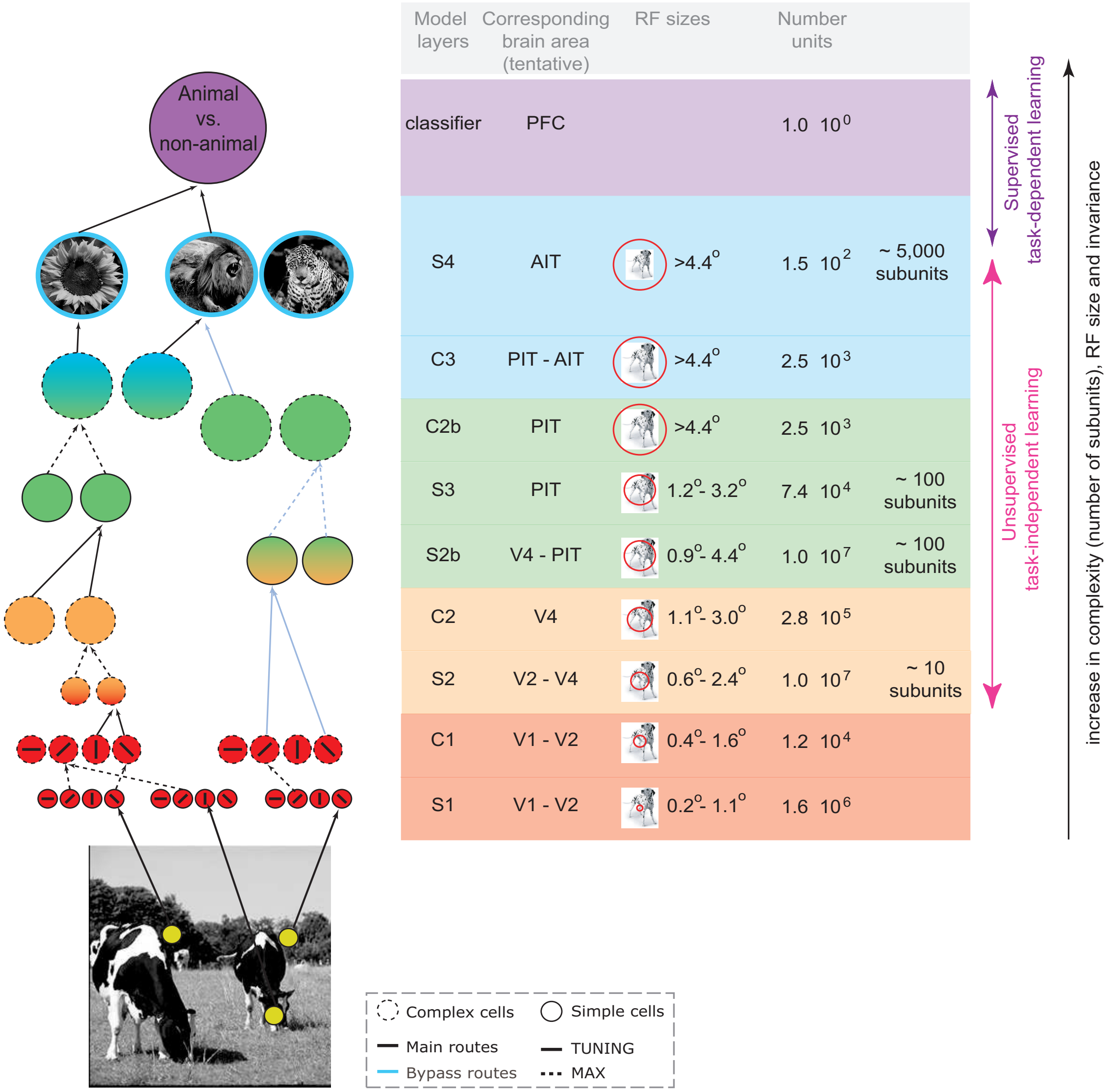
One of the first feedforward models, Fukushima's Neocognitron [Fukushima, 1980], followed the basic Hubel & Wiesel hierarchy [Hubel and Wiesel, 1968] in a computer vision system. Building upon several conceptual proposals [Perrett and Oram, 1993; Wallis and Rolls, 1997; Mel, 1997; Hochstein and Ahissar, 2002; Ullman et al., 2002; Thorpe, 2002; Amit and Mascaro, 2003; Wersing and Koerner, 2003], we developed [Riesenhuber and Poggio, 1999a; Serre et al., 2002; Giese and Poggio, 2003; Serre et al., 2005a] a similar computational theory that attempts to quantitatively account for a host of recent anatomical and physiological data. The model [Riesenhuber and Poggio, 1999a; Serre et al., 2002, 2005a] shown in Fig. 5-1 (see Methods) is qualitatively and quantitatively consistent with (and in some cases actually predicts) several properties of cells in V1 [Lampl et al., 2004], V2, V4 (see Chapter 3 and [Serre et al., 2005a; Gawne, 2000]) and IT [Riesenhuber and Poggio, 1999a] as well as fMRI and psychophysical data [Riesenhuber et al., 2004]. Plausible biophysical circuits may implement the two key operations (see Chapter 2 assumed by the theory within the time constraints of the experimental data [Perrett et al., 1992; Hung et al., 2005]).

The main extension with respect to the original model [Riesenhuber and Poggio, 1999a] is an unsupervised learning of the tuning of each unit at the S_2, S_{2b} and S_3 levels (possibly corresponding to V4 and PIT, see Fig. 5-1 and Methods) on a set of natural images unrelated to the task. In the present model, units (of the S type) become tuned to the neural activity induced by natural images within their receptive fields. We conjecture from our simulations that the resulting large number of tuned units constitutes a universal and redundant dictionary of features [Ullman et al., 2002], which is invariant (to some extent) to translation and scale and can support the recognition of many different object categories (such as animals, cars and faces, see Chapter 4 and [Serre et al., 2005b]). When tested on real-world natural images, the model competes with and sometimes even outperforms state-of-the-art computer vision systems on several categorization tasks [Serre et al., 2005a,b] (see also Chapter 4). This is quite surprising, given the many specific biological constraints that the theory satisfies.

It is therefore natural to ask whether any such feedforward model may be able to duplicate human-level performance in natural, complex recognition tasks. Normal, everyday vision includes top-down effects which must be mediated by the extensive anatomical back-projections found throughout visual cortex [Bullier, 2001]. Back-projections may effectively control the "programs" and circuits, for instance in PFC, that read out in a task-dependent way the information from lower visual areas (*e.g., is the object in the scene an animal? how big is it?*) [Hung et al., 2005]. They could in addition influence areas lower than IT during or before the task, for instance by modulating connections. The key claim of feedforward models, such as the one presented here, is that the first $150\ ms$ of visual perception do not involve significant feedback dynamics.

Just like an experimental test of Newton's second law requires choosing a situation in which friction is negligible, we looked for an experimental paradigm in which recognition has to be fast and cortical back-projections are likely to be inactive. The paradigm we use to compare human performance to that of a feedforward model of visual processing is ultra-rapid object categorization. The task is the classical animal *vs.* non-animal recognition task [Thorpe et al., 1996; Thorpe and Fabre-Thorpe, 2001; VanRullen and Koch, 2003; Rousselet et al., 2003; Bacon-Mace et al., 2005]. Animals in natural scenes constitute a challenging class of stimuli due to large variations in shape, pose, size, texture, and position in the scene.

We used a backward masking protocol (1/f noise image with a duration of $80\ ms$, see Fig. 5-2a). Previous studies have suggested that a backward mask can interrupt visual processing [Kovács et al., 1995; Rolls et al., 1999; Keysers et al., 2001] and block back-projections [Bacon-Mace et al., 2005]. To vary the difficulty of the task and prevent human observers and the model from relying on low-level cues, we used four sets of balanced image categories (150 animals and 150 matching distractors), each corresponding to a particular viewing-distance from the camera, from an animal head to a small animal or groups of animals in cluttered natural backgrounds (*i.e., head*, *close-body*, *medium-body* and *far-body* categories, see Fig. 5-2b and Supp. Info.).



Model layers	Corresponding brain area (tentative)	RF sizes	Number units	
classifier	PFC		$1.0\ 10^0$	
S4	AIT	>4.4°	$1.5\ 10^2$	~ 5,000 subunits
C3	PIT - AIT	>4.4°	$2.5\ 10^3$	
C2b	PIT	>4.4°	$2.5\ 10^3$	
S3	PIT	1.2°- 3.2°	$7.4\ 10^4$	~ 100 subunits
S2b	V4 - PIT	0.9°- 4.4°	$1.0\ 10^7$	~ 100 subunits
C2	V4	1.1°- 3.0°	$2.8\ 10^5$	
S2	V2 - V4	0.6°- 2.4°	$1.0\ 10^7$	~ 10 subunits
C1	V1 - V2	0.4°- 1.6°	$1.2\ 10^4$	
S1	V1 - V2	0.2°- 1.1°	$1.6\ 10^6$	

Animal vs. non-animal
Supervised task-dependent learning
Unsupervised task-independent learning
increase in complexity (number of subunits), RF size and invariance
Complex cells
Simple cells
Main routes
Bypass routes
TUNING
MAX

Figure 5-1: Schematic of the model implementation used in the comparison with human-observers on the animal *vs.* non-animal categorization task. The theory assumes that one of the main functions of the ventral stream is to achieve a trade-off between selectivity and invariance. As in [Riesenhuber and Poggio, 1999a], stages of *simple* S units with Gaussian tuning (plain circles and arrows), are interleaved with layers of *complex* C units (dotted circles and arrows), which perform a max operation on their inputs and provide invariance to position and scale (pooling over scales is not shown in the figure). The major extension in this model relative to [Riesenhuber and Poggio, 1999a] is that unsupervised learning, on a set of natural images unrelated to the task, determines the tuning (*e.g.*, the synaptic weights) of the simple units in the S_2 and S_3 layers (corresponding to V4 and PIT, respectively). Learning of the synaptic weights from S_4 to the top classification units is the only task-dependent, supervised learning stage in this architecture. The total number of units in the model is in the order of 10^7. Colors indicate the correspondence between model layers and cortical areas. The table on the right provides a summary of the main properties of the units at the different levels of the model. The diagram on the left is modified from Van Essen & Ungerleider [Gross, 1998] (with permission by the authors).

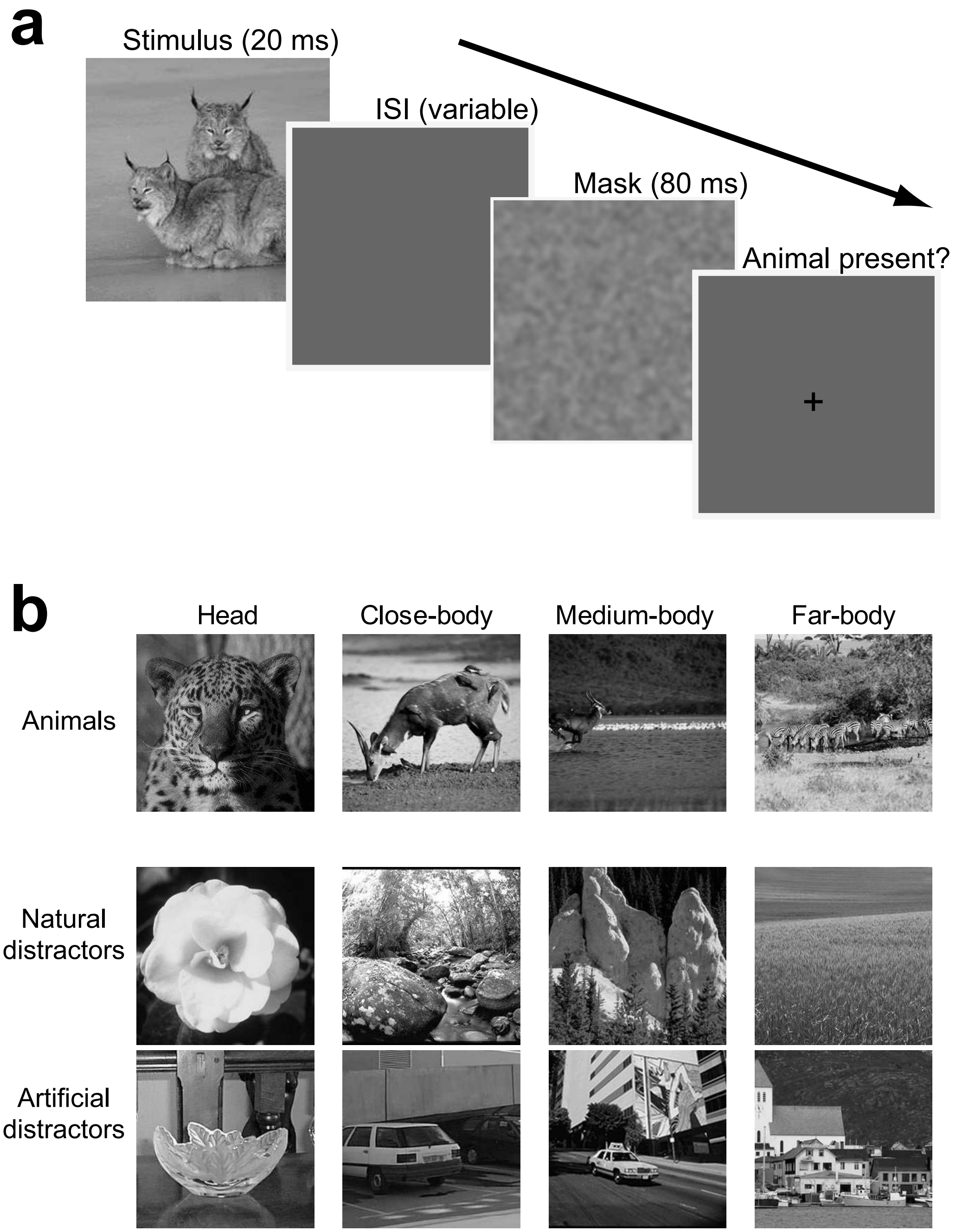


Figure 5-2: a) Schematic of the task. A stimulus (gray-level image) is flashed for $20\ ms$, followed by a blank screen for a variable duration denoted ISI (inter-stimulus interval) and followed by a mask for $80\ ms$. We tested four conditions: immediate-mask, $30\ ms$ ISI, $60\ ms$ ISI and no-mask. Subjects ended the trial with a yes/no answer by pressing one of two keys. **b)** The four (balanced) classes of stimuli. Animal images (a subset of the image database used in [Thorpe et al., 1996]) were manually arranged into four groups (150 images each) based on the animal-distance from the camera: head (close-up), close-body (animal body occupying the whole image), medium-body (animal in scene context) and far-body (small animal or groups of animals). Each of the four classes corresponds to different animal sizes and modulates the task difficulty (see Fig. 5-3). A set of matching distractors (300 each from natural and artificial scenes, see Supp. Info.) was selected, so as to prevent human observers and the computational model from relying on low-level cues.

Before the model can be tested on the animal *vs.* non-animal categorization task, it has to be trained. The only task-specific training required involves the circuits at the top level in the model, *i.e.*, the linear classifier (possibly at a level such as PFC) that "looks" at the activity of several hundred S_4 units [Hung et al., 2005]. Such classifier is trained on a specific task (*i.e.*, animal *vs.* non-animal) in a supervised way (see Methods and Supp. Info.). This stage requires a relatively small number of examples (~ 100). The classifier was trained using n random splits on the entire database of images. In a given run, half the images were selected at random for training and the other half was used for testing the model (see Supp. Info.).

In the present version of the model, processing by the units (the nodes of the graph in Fig. 5-1) is approximated as essentially instantaneous (see however possible microcircuits involved in the tuning and max operation in [Serre et al., 2005a]). All the processing time would be taken by synaptic latencies and conduction delays (see Supp. Info.). The model was compared to human observers in three different experiments.

B Experiment 1

In experiment 1, we replicated previous psychophysical results [Bacon-Mace et al., 2005] to test the influence of the mask on visual processing with four experimental conditions, *i.e.*, when the mask followed the target image a) without any delay (*immediate-mask* condition), b) with a short inter-stimulus interval of $30\ ms$ (*30 ms ISI*), c) with an ISI of $60\ ms$ or d) never (*no-mask* condition). For all four conditions, the target presentation was fixed to $20\ ms$. The performance in immediate- and no-mask conditions establishes lower and upper bounds on human performance (in the absence of eye movements). A comparison between the performance of human observers ($n = 21$) and the feedforward model is shown in Fig. 5-3.

We found that the accuracy (hits) of the human observers was well within the range of data previously obtained with go/no-go tasks [Thorpe et al., 1996; VanRullen and Koch, 2003; Bacon-Mace et al., 2005]. The subjects' level of performance reached a ceiling in the $60\ ms$ ISI condition (except when the animal was camouflaged in the scene, *i.e.*, far-body group). As expected, the delay between the stimulus and the mask onset (*i.e.*, the ISI) modulates the level of performance of the observers, improving from bare recognition in the immediate-mask condition to ceiling in the no-mask condition.

The shaded area in Fig. 5-3c defines a range of likely admissible detection performance that would correspond to the human visual system operating in its feedforward mode (see Supp. Info.). The model performance is well within this area and predicts human-level performance between the $30\ ms$ ISI and $60\ ms$ ISI conditions. The implication is that, for this range of ISI, the back-projections do not play a significant role and that the model may indeed provide a satisfactory description of the feedforward path (see Supp. Info.).

C Experiment 2

In experiment 2, we further refined the comparison between the model and human observers by testing subjects ($n = 24$) on a single mask condition ($30\ ms$ ISI). To take into account responses to both target and distractor stimuli, we report here a sensitivity measure from signal detection theory [Macmillan and Creelman, 1991] called d', that is the standardized difference between the means of the hit and false-alarm distributions of each observer (error rates and hits would give similar results, see Supp. Info.).

As shown in Fig. 5-4a, human observers behave similarly to the model: for all four animal categories, their levels of performance do not show significant differences (with overall correct 80% for human observers and 82% for the model) and they both exhibit a similar trend on the four groups (close-body being the simplest and far-body the most difficult). The performance of the model is remarkable, given the comparatively lower performance of other computational systems that have been previously compared to human observers on rapid categorization tasks and that rely on low-level cues.

The benchmark computer vision systems (see Supp. Info.) were Torralba & Oliva's global features [Torralba and Oliva, 2003] (75% correct) and Malik and colleagues' textons [Renninger and Malik, 2004] (62% correct). Lower levels in the hierarchical architecture of the model did also have a lower performance (see Fig. 5-4a, Supp. Info. and [Torralba and Oliva, 2003]).

We also looked at the agreement between human observers and the model on individual images (see Fig. 5-4b). For each image in the database, we computed the percentage of subjects (right number above each thumbnail in Fig. 5-4b) who classified it as an animal (irrespective of whether the image contains an animal or not). For the model, we com-

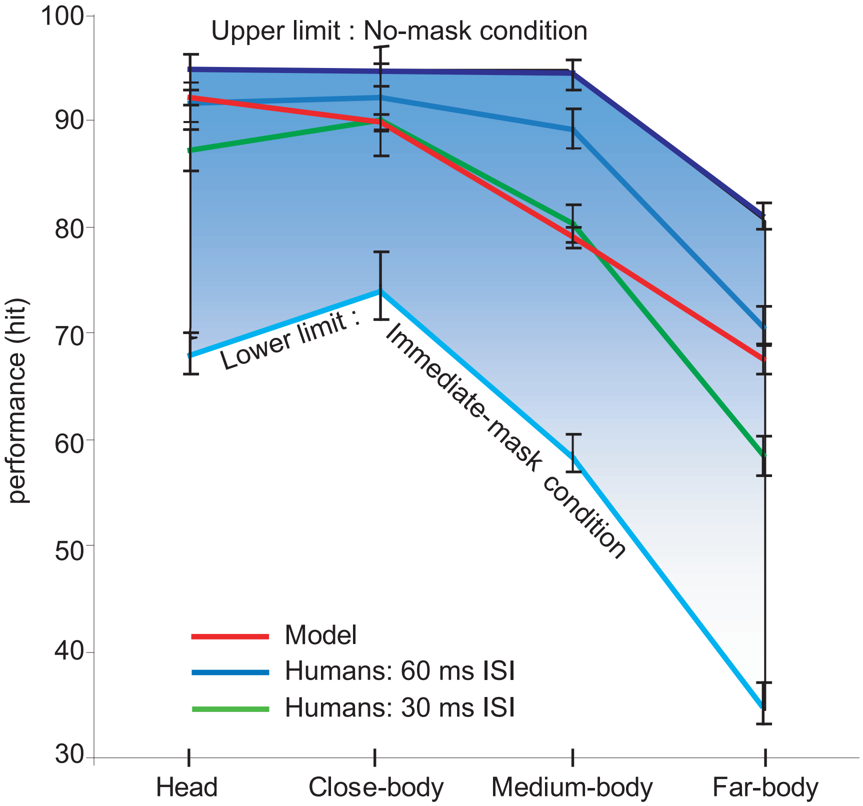


Figure 5-3: Experiment 1: Comparison between the model and human observers with different mask conditions. Model *vs.* human level accuracy measured by hits, *i.e.*, the percentage of animals correctly detected, for comparison with results using go/no-go tasks [Thorpe et al., 1996; VanRullen and Koch, 2003; Bacon-Mace et al., 2005] (see Supp. Info. for error rates). The upper and lower bounds on human-level performance ($n = 21$) are given by the no-mask condition (from 95% to 81%) and the immediate-mask condition (from 74% to 35%) respectively. The average accuracy of human observers for the conditions with immediate-mask, 30 ms ISI, 60 ms ISI and no-mask conditions were $59\%, 79\%, 86\% and 91\%$ respectively - all significantly above chance (t-test, $p < 0.01$) – compared to 82% for the model. The accuracy for all conditions is comparable to previously published results in go/no-go tasks [Thorpe et al., 1996; VanRullen and Koch, 2003; Bacon-Mace et al., 2005]. For human observers, the false-alarm rate does not vary significantly with the various backward masking conditions ($16\%, 16\%, 16\%$ and 14%). The model matches human observers for ISIs between 30 ms and 60 ms. Error bars indicate the standard error and are not directly comparable for the model (computed over N random runs, see Supp. Info.) and for humans (computed over n observers).

puted the percentage of times the model (left number) classified each image as an animal for each of the random runs ($n = 20$). A percentage of 100% (50%) means that all (half) the observers (either human observers or random runs of the model) classified this image as an animal. The overall correlation on the percentages for the model and for human observers was $0.71, 0.84, 0.71$ and 0.60 for heads, close-body, medium-body and far-body respectively (all values were statistically significant, $p < 0.01$).

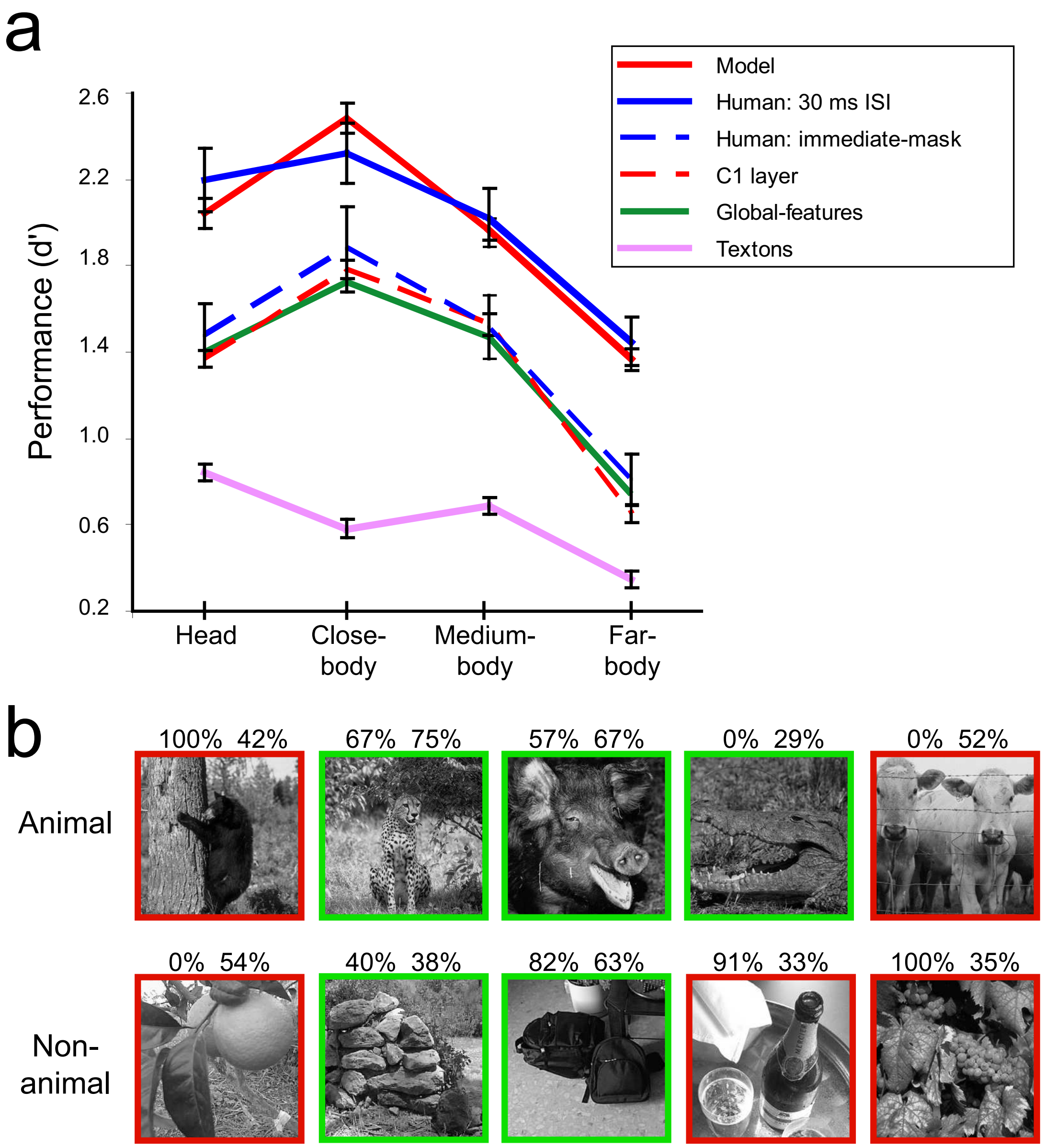


Figure 5-4: Experiment 2: Detailed comparison between the model, other computational benchmarks and human observers. a) Model *vs.* other computational benchmarks *vs.* human-level accuracy. To account for both hit and false-alarm rates, we here report the d' sensitivity measure [Macmillan and Creelman, 1991] (see text). Both the model and human observers performed best on close-body views and worst on far-body views. Other computational benchmarks of object recognition that rely on a combination of low-level features [Torralba and Oliva, 2003; Renninger and Malik, 2004] were run on the same animal database (with the same training and test sets as for the model) and showed lower categorization performance (see text). b) Comparison between human observers and the model on individual images. From left to right are representative images for which the model went from being correct to incorrect. The percentages above each thumbnail correspond to the number of times the image was classified as animal by the model (left) or by human observers (right, see text for details). Green (red) bounding boxes correspond to images for which human observers and the model agree (disagree).

D Experiment 3

Finally, in experiment 3, we measured the effect of image rotation (90^o and 180^o). Recent behavioral studies [Rousselet et al., 2003; Guyonneau et al., 2005] suggested that the animal categorization task can be performed at different image orientations, thus providing an interesting test for the model. As shown in Fig. 5-5, the level of performance of both human observers (left) and the model (right) is quite robust to image rotation (except for the far-body condition for which the prominent scene background is likely to influence performance). The accuracy measures obtained for human observers and the model (see Supp. Info.) are compatible with previous results by Thorpe and collaborators [Rousselet et al., 2003; Guyonneau et al., 2005]. The robustness of the model is particularly remarkable as it was not re-trained before being tested on the rotated images (it is unlikely that human subjects had extensive experience with rotated images of animals). The fact that a feedforward model - faithful to the physiology and anatomy of visual cortex - achieves a level of accuracy comparable to humans on a difficult recognition task raises an intriguing question with potentially rich implications for research in different domains of visual sciences: what are the really difficult purely visual recognition tasks that need feedback and the involvement of back-projections?

E Methods

Here we give a brief overview of the model implementation and learning techniques used. Details about the model can be found in Chapter 2. Details on the human psychophysics experiments can be found in Supp. Info.

E.1 Model architecture

The first stage of simple units (S_1) which corresponds to the classical simple cells of Hubel & Wiesel [Hubel and Wiesel, 1968], represents the result of a first tuning operation: Each S_1 unit receives LGN-like inputs and is tuned to an oriented bar with a Gaussian-like profile. Each of the complex units (C_1) in the second layer pools the outputs of a group of neighboring simple units in the first layer. These units are at slightly different positions and sizes but have the same preferred orientation. The pooling is performed by a max

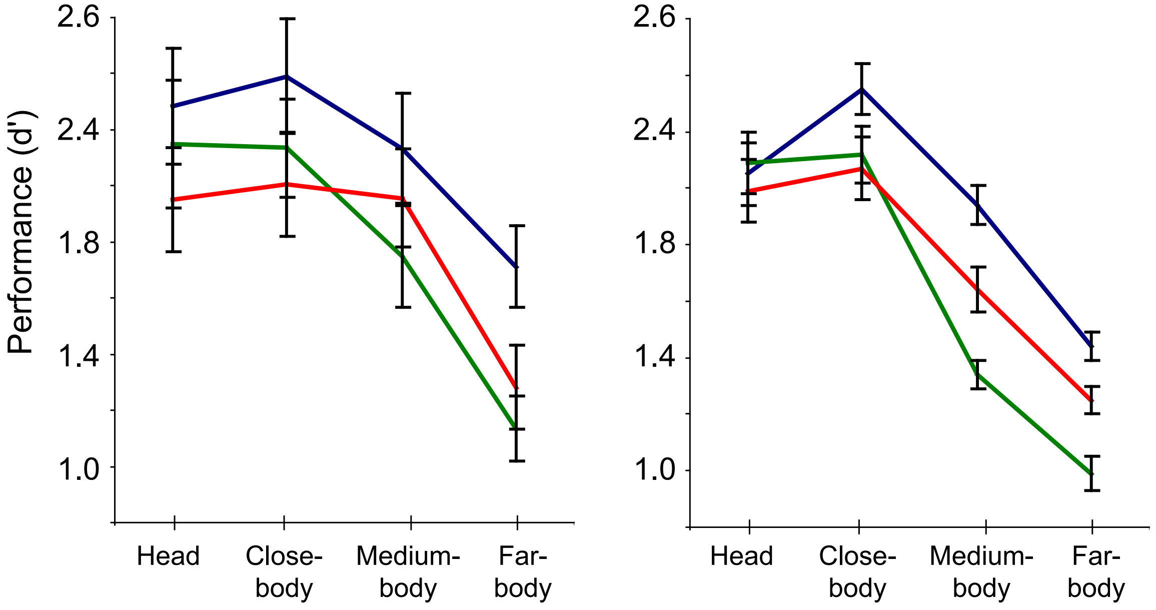


Figure 5-5: Experiment 3: Comparison between the model and human observers on rotated images. We compared the performance (d') obtained in three experimental conditions: upright, 90^o rotation and inverted (180^o) for human observers (left) and the model (right). Both human observers and the model were robust to image rotations (except for the far-body condition) and exhibited similar patterns of performance.

operation such that the activity of the complex pooling unit is equal to the activity of the strongest input.

At the next layer, each simple (S_2) unit pools several complex (C_1) units - with weights dictated by the unsupervised learning stage - with different selectivities according to a Gaussian tuning function, thus yielding selectivity to more complex patterns. Simple units in higher layers (S_3 and S_4) combine more and more complex features with a Gaussian tuning function, while the complex units (C_2 and C_3) pool their outputs through a max function providing increasing invariance to position and scale. In the model, the two layers alternate (though levels could conceivably be skipped, it is likely that only units of the S type follow each other above C_3).

Here we use a multivariate Gaussian for the tuning operation (see Eq. 1.2. The weight vector $\mathbf{w}$ is learned with no supervision from natural images (see below). A complex unit activity is given by a max operator (see Eq. 1.2). Despite the fact that a max operation seems very different from a Gaussian tuning, they can both be implemented in terms of biologically plausible normalized scalar product operations with a gain control circuit.

Learning a universal dictionary of shape-tuned units in the model. Each unit in the simple layers (S_2, S_{2b} and S_3 sequentially) becomes tuned by exposing the model to a set of 1,000 natural images. For each image presentation, units become tuned to the pattern of activity of their afferents (see Supp. Info.). This learning stage is similar to "imprinting" and could possibly be mediated by a mechanism of the LTP type. In the model the learning stage corresponds to setting each $\mathbf{w}$ to the pattern of pre-synaptic activity.

E.2 Classifier from IT to PFC

The linear classifier from IT to PFC used in the simulations corresponds to a supervised learning stage with the form: (3) where characterizes the response to the input image $\mathbf{x}$ of the i^{th} S_4 unit tuned to the training example $\mathbf{x^i}$ (animal or non-animal) and $\mathbf{c}$ is the vector of synaptic weights from IT to PFC. The superscript i indicates the index of the image in the training set and the subscript j indicates the index of the pre-synaptic unit. Since the S_4 units (corresponding to the view-tuned units in IT [Logothetis et al., 1995]) are like Gaussian radial basis functions (RBFs), the part of the network in Fig. 5-1 comprising the inputs to the S_4 units up to PFC can be regarded as an RBF network (see Supp. Info.). Supervised learning at this stage involves adjusting the synaptic weights c so as to minimize a (regularized) error on the training set [Poggio and Bizzi, 2004] (see Supp. Info.).

F Supplementary Information

F.1 Supplementary Methods

Categorization by the human observers

For all three experiments, participants gave a written informed consent. All participants were between 18 and 35 years old, with $n = 21, 24$ and 14, in experiments 1, 2 and 3 respectively. There was approximately the same number of male and female observers in each experiment and none participated in more than one of the three experiments. Participants were seated in a dark room, $0.5\ m$ away from a computer screen, connected to a computer (Intel Pentium© IV processor, 1 GB RAM, 2.4 GHz). The monitor refresh rate was 100 Hz allowing stimuli to be displayed with a frame-duration of $10\ ms$ and a resolution of 1024×768. We used the Matlab© (Mathworks Inc, Natick, MA) software with the psychophysics toolbox [Brainard, 1997; Pelli, 1997] to precisely time the stimulus presentations. In all experiments, the image duration was $20\ ms$. In experiment 1, the mask appeared after an inter-stimulus interval (ISI) of $0\ ms$ (corresponding to a Stimulus Onset Asynchrony – SOA – of $20\ ms$), $30\ ms$ ($SOA = 50ms$), $60\ ms$ ($SOA = 80ms$), or infinite (*i.e.,* never appeared). In experiments 2 and 3, we tested a fixed ISI of $30\ ms$ ($SOA = 50ms$). The mask following the picture was a (1/f) random noise mask, generated by filtering random noise through a Gaussian filter. The stimuli were presented in the center of the screen (256×256 pixels, gray-level images). All images had a mean average luminance of 128 with pixel intensities ranging from 0 to 255. The 600 animal stimuli were grouped into four categories with 150 exemplars in each, *i.e., head*, *close-body*, *medium-body* and *far-body*. A set of distractors with matching mean distance from the camera (300 from natural and 300 from artificial scenes) was selected from a database of annotated mean depth images [Torralba and Oliva, 2002]. We selected images with a mean distance from the camera below 1 m for head, between $5\ m$ and $20\ m$ for close-body, between $50\ m$ and $100\ m$ for medium-body as well as above 100 m and panoramic views for far-body. The 1,200 image stimuli (600 animals and 600 distractors) were presented in random order and divided into 10 blocks of 120 images each. Participants were asked to answer as fast and as accurately as possible if the image contained an animal, by pressing a *yes* or *no* key on a computer keyboard. They were randomly asked to use their left or right hand for yes *vs.* no answers. Each experiment took about thirty minutes to perform.

Task-independent unsupervised learning in the model

Here we used an extended version [Serre et al., 2002, 2005a] of the original model [Riesenhuber and Poggio, 1999a] that relies on a simple learning rule to determine the tuning of the S units from visual experience. In the original implementation of the model [Riesenhuber and Poggio, 1999a] learning only occurred in the top-most layers (*i.e.,* the units that correspond to the view-tuned units in AIT [Logothetis et al., 1995] and the task-specific circuits from IT to PFC [Freedman et al., 2001]). In this initial simple version it was possible to manually tune units in intermediate layers (simple 2×2 combinations of 4 orientations, see [Riesenhuber and Poggio, 1999a]) to be selective for the target object. It turns out that the extended version with learning at all stages is more faithful to the physiology data and performs significantly better in recognizing real-world images (such as faces with different illuminations, background, expression, *etc*) [Serre et al., 2002; Louie, 2003; Serre et al., 2005b,a].

During training, the model was exposed to a set of natural images $(1,000)$ collected from the web (including landscapes, street scenes, animals) and unrelated to the categorization task. For each image presentation, units became tuned to the pattern of activity of their afferents. This was done for each layer sequentially, starting from bottom to top (*i.e.*, S_2 and S_{2b} first then S_3). This can be regarded as an imprinting process in which each S unit (*e.g.*, , S_2 unit) stored in its synaptic weights the specific pattern of activity from its afferents (*e.g.*, C_1 units) in response to the part of the natural image that fell within its receptive field.

In the Gaussian approximation used here (see Methods) this was done by setting w to the pattern of pre-synaptic activity. A biologically plausible version of this rule could involve mechanisms such as LTP. The image patch that fell within the receptive field of a unit became its preferred stimulus with a bell-shape tuning profile. We assumed that the images move (shifting and looming) so that each type of S unit was replicated across the visual field. The tuning of units from S_1 to S_3 is fixed after this development-like stage. Afterward, only the task-specific circuits from IT to PFC required learning for the recognition of specific objects and object categories.

Task-dependent supervised learning and categorization by the model

We trained the classifier on a set of training examples as $(\mathbf{x^i}, y^i)$ pairs, where $\mathbf{x^i}$ denotes the i^{th} image in the training set and y^i its associated label (animal or non-animal). To train the classifier that corresponds to the task-specific circuits from IT to PFC, we used a random-split procedure which has been shown to give good estimates of a classifier expected error [Devroye et al., 1996]. We performed $n = 20$ random runs. In each run, half of the 1,200 image examples from the database of stimuli in experiment 1, 2 and 3 was used for training the model and the remaining half for testing it. For a test image $\mathbf{x}$, the classifier response is given by Eq. 2.1.

The model performance reported in experiment 1, 2, and 3 was averaged over these n random runs. Note that the error bars for the model in Fig. 5-3, 5-4, 5-5 correspond to the standard errors computed over these $n = 20$ random runs. Error bars are therefore not directly comparable with those for human observers. In a separate experiment we trained four classifiers, one for each animal type (see Fig. 5-3), and further aggregated their outputs for the animal *vs.* non-animal classification with similar results. It is possible that better results could be obtained by training separate classifiers for different animal species and then aggregating their outputs. In general, increasing the set of supervised examples should improve the performance on the task.

Categorization by the benchmark computer systems

Global (context) features [Torralba and Oliva, 2003] were computed by convolving each image in the database with a filter pyramid (24 Gabor filters covering several orientations and scales) and further down-sampling to produce the resulting $4 \times 4 \times 24$ image (4×4 is the number of samples used for each filter in this low-resolution representation). The dimensionality of each of these $4 \times 4 \times 24$ vectors was further reduced by applying principal component analysis [Torralba and Oliva, 2003] producing, for each image in the database, a feature vector $\mathbf{x}$ that provides a low-resolution encoding of the distribution of orientations and scales across the entire image. The system performance was evaluated using $n = 10$ random splits (see [Oliva and Torralba, In press]). In each run half the images were used to train a linear classifier on the feature vector $\mathbf{x}$ and the remaining half to evaluate its performance.

Textons features [Renninger and Malik, 2004]. The software for the texture descriptors called textons [Renninger and Malik, 2004] was kindly provided by Stan Bileschi at CBCL (see paper by Bileschi & Wolf [Bileschi and Wolf, 2005]) directly as input to a classifier. For each image in the database, a feature vector $\mathbf{x}$ was computed by concatenating the response of a fixed subset of $1,500$ C_1 model units. The system performance was evaluated using ($n = 20$) random runs as for the model. In each run half the images were used to train a linear SVM classifier on the feature vector $\mathbf{x}$ and the remaining half to evaluate its performance. We found that the C_1 layer responses yield a performance which is very similar to the performance of humans on the immediate-mask condition ($ISI = 0$).

It is interesting to point out that the level of performance of the C_1 layer is very similar to the level of performance of the global context features of Torralba & Oliva. Because the computational benchmarks rely on low-level features, it is not surprising that that they perform worse than the feedforward model on a high-level recognition task such as animal *vs.* non-animal categorization. This suggests the need for a representation based on units with different levels of complexity and invariance as in the architecture of 5-1. An independent study [Hung et al., 2005; Serre et al., 2005a] found a gradual improvement (using layers in the model from bottom to top) in reading out several object categories (at different positions and scales) from various model layers.

F.2 On Interrupting Back-Projections with the Mask

There is much debate about the effect of a mask – as used in the psychophysics described here – on visual processing. A well accepted theory is the "interruption theory" that has been in fact corroborated by physiological studies [Rolls and Tovee, 1994; Tovee, 1994; Kovács et al., 1995; Rolls et al., 1999; Keysers et al., 2001] (see also [Lamme and Roelfsema, 2000]). The assumption is that the visual system processes stimuli sequentially (in a pipeline-like architecture): when a new stimulus (the mask) is piped in, it interrupts the processing of the previous stimulus (the target image).

Here we would like to try to isolate a purely feedforward sweep from further recurrent processing [Lamme and Roelfsema, 2000]. Whether or not the back-projections may participate in the overall processing and contribute to the final performance is determined by the delay between the stimulus and the mask, *i.e.*, the SOA. If the delay Δ taken by the visual signal to travel from stage A to stage B and back to stage A is longer than the SOA, this back-projection will not influence the processing in the visual system as it will be interrupted before.

Based on estimates of conduction delays (see Fig. 5-6), extrapolated from monkey [Nowak and Bullier, 1997; Thorpe and Fabre-Thorpe, 2001] to human [Thorpe, pers. comm.], we think that in all our experiments, a SOA of $50\ ms$ is likely to be the longest SOA before significant feedback loops become active[2], for instance, between IT and V4 (see Fig. 5-6, orange arrows, $\Delta \sim 40 - 60\ ms$). Importantly such an SOA should exclude major top-down effects, for instance between IT and V1 ($\Delta \sim 80 - 120\ ms$), while leaving enough time for signal integration at the neural level.[3]

This estimate seems in good agreement with results from a Transcranial Magnetic Stimulation (TMS) experiment [Corthout et al., 1999] that has shown a disruption of the feedforward sweep [Lamme and Roelfsema, 2000] for pulses applied between $30\ ms$ and $50\ ms$ after stimulus onset.[4] It is thus quite interesting that the model matches human perfor-

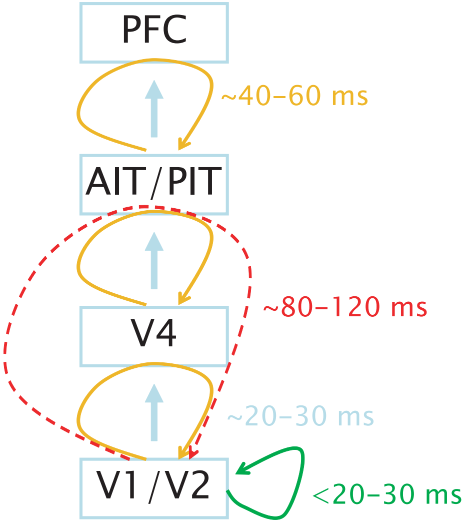


Figure 5-6: Estimate of the timing of feedback loops in the ventral stream of primate visual cortex (based on [Nowak and Bullier, 1997; Thorpe and Fabre-Thorpe, 2001]. We assume that typical latencies from one stage to the next is $\sim 10\ ms$ and that feedforward and back-projections have similar conduction times [Nowak and Bullier, 1997]. The first number corresponds to latencies for monkeys and is assumed to constitute a lower bound on the latencies for humans. The second number corresponds to an additional 50% and is assumed to constitute a "typical" number for humans [Thorpe, pers. comm.].

mance almost exactly for an SOA of $50\ ms$, but underperforms it for longer $SOAs$. One of the possible explanations is that this is due to back-projections which are not included in the present, purely feedforward model of Fig. 1.

F.3 Supplementary Data

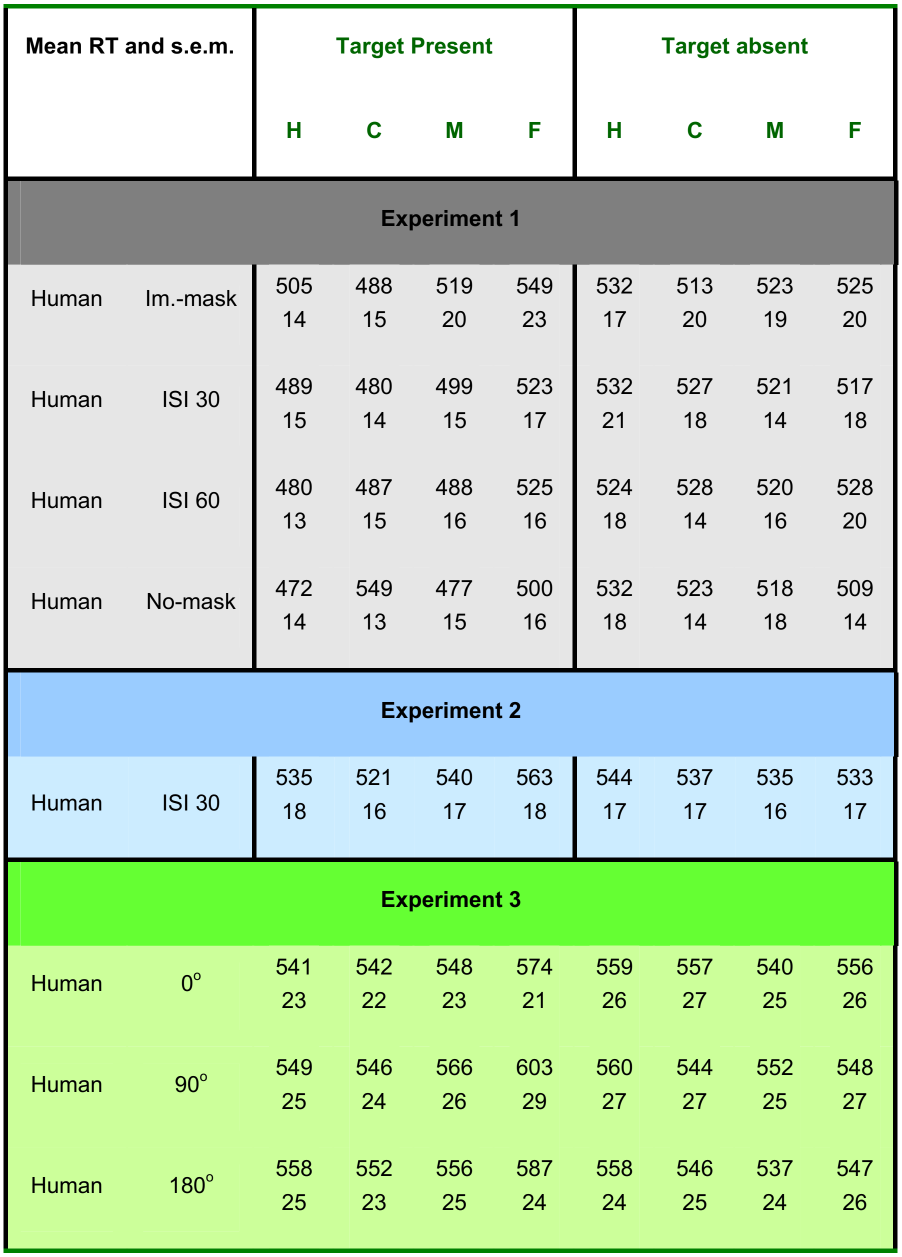
Tables 5.1 summarizes the mean and standard error of the reaction times for human observers and Tables 5.2 the 10^{th} percentile. Tables 5.3 and 5.4 summarizes the main accuracy measurements or both human observers and the model.

Notes

[1] Interestingly, consistent with the model described here, a recent RSVP study showed that during a detection task, while observers were able to correctly detect the target, they were, however, unable to accurately locate the target [Karla and Treisman, 2005].

[2] Note that for such SOA, local feedback loops green arrows in Fig. 5-6) are likely to be already active ($\Delta < 20 - 30\ ms$), see [Knierim and van Essen, 1992; Zhou et al., 2000].

[3] The mask is likely to interrupt the maintained response of IT neurons but not to alter their initial selective response [Kovács et al., 1995; Rolls et al., 1999]. According to an independent study [Hung et al., 2005] this would provide significantly more time than needed

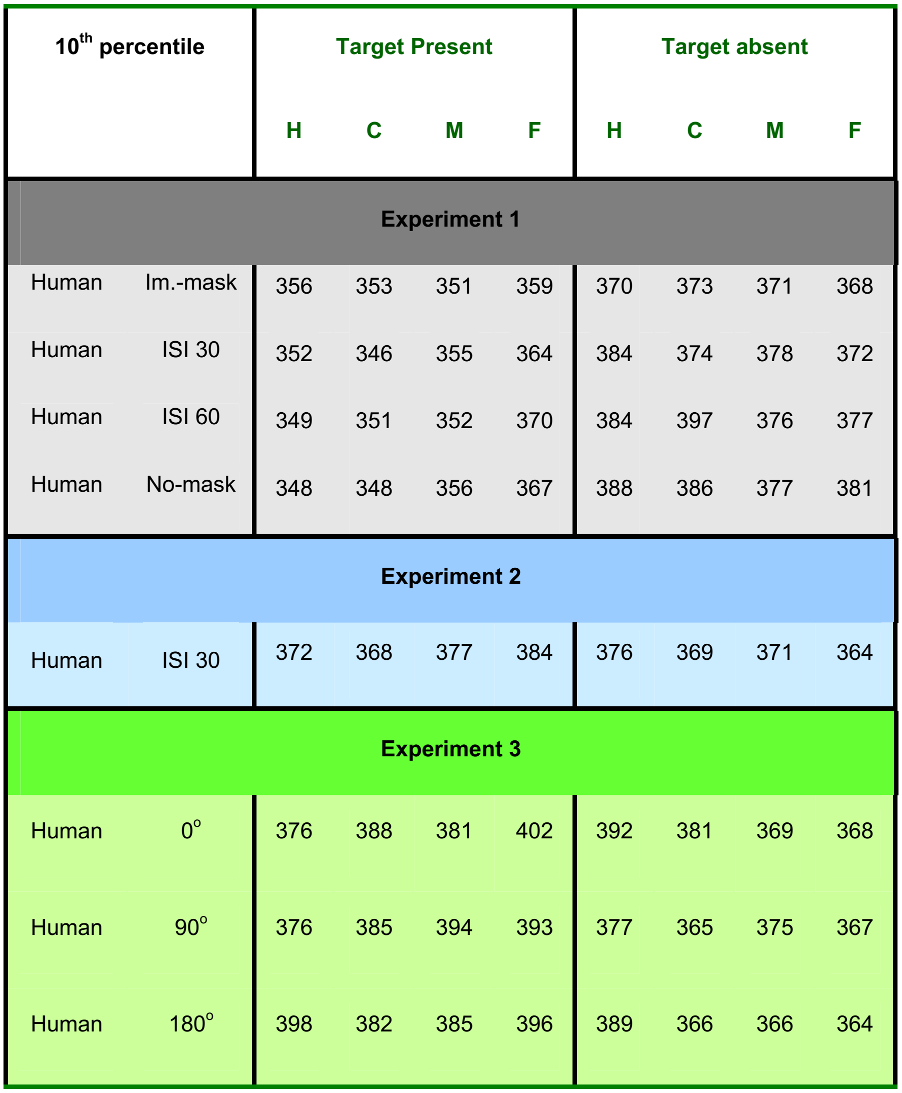


Mean RT and s.e.m.		Target Present				Target absent			
		H	C	M	F	H	C	M	F
Experiment 1									
Human	Im.-mask	505 14	488 15	519 20	549 23	532 17	513 20	523 19	525 20
Human	ISI 30	489 15	480 14	499 15	523 17	532 21	527 18	521 14	517 18
Human	ISI 60	480 13	487 15	488 16	525 16	524 18	528 14	520 16	528 20
Human	No-mask	472 14	549 13	477 15	500 16	532 18	523 14	518 18	509 14
Experiment 2									
Human	ISI 30	535 18	521 16	540 17	563 18	544 17	537 17	535 16	533 17
Experiment 3									
Human	0°	541 23	542 22	548 23	574 21	559 26	557 27	540 25	556 26
Human	90°	549 25	546 24	566 26	603 29	560 27	544 27	552 25	548 27
Human	180°	558 25	552 23	556 25	587 24	558 24	546 25	537 24	547 26

Table 5.1: Summary of mean reaction times *(mean RT)* and standard error mean *(s.e.m.)* for human observers on correct responses (in ms).

($\gg 12.5\ ms$) to permit robust recognition in “reading out” from monkey IT neurons.

[4] The same experiment [Corthout et al., 1999] also demonstrated blockade of perception by pulses applied between $80 - 120\ ms$, presumably corresponding to recurrent processing [Lamme and Roelfsema, 2000] by the back-projections.

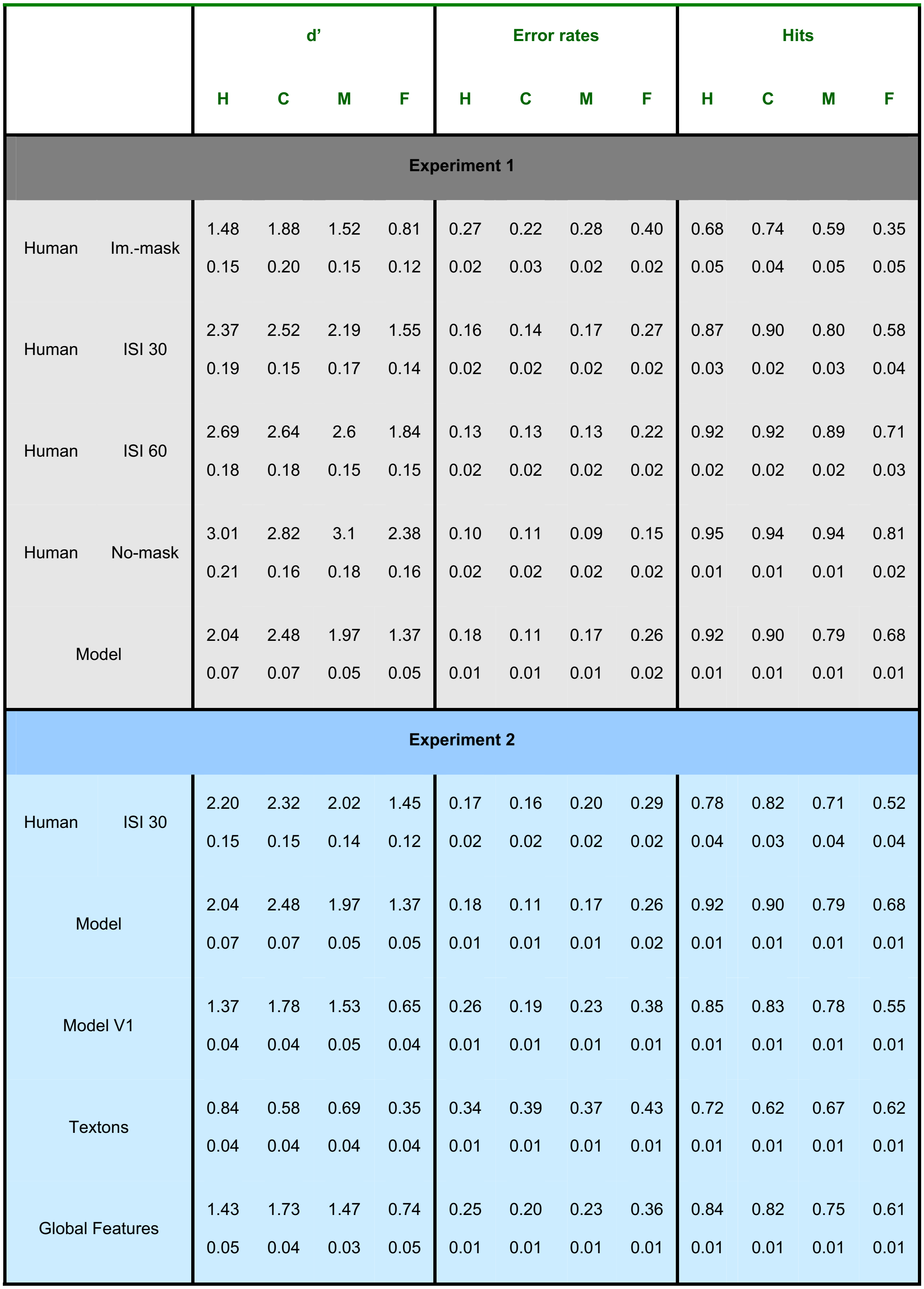


10th percentile		Target Present				Target absent			
		H	C	M	F	H	C	M	F
Experiment 1									
Human	Im.-mask	356	353	351	359	370	373	371	368
Human	ISI 30	352	346	355	364	384	374	378	372
Human	ISI 60	349	351	352	370	384	397	376	377
Human	No-mask	348	348	356	367	388	386	377	381
Experiment 2									
Human	ISI 30	372	368	377	384	376	369	371	364
Experiment 3									
Human	0°	376	388	381	402	392	381	369	368
Human	90°	376	385	394	393	377	365	375	367
Human	180°	398	382	385	396	389	366	366	364

Table 5.2: Summary of reaction times (10^{th} percentile) for human observers on correct responses (ms).

Acknowledgements

This work was done in collaboration with Aude Oliva. We would like to thank Stan Bileschi for re-implementing and running the texton algorithm. We are very grateful to Sanmay Das, Robert Desimone, Jim DiCarlo, David Ferster, Martin Giese, Michelle Greene, Christof Koch, Ethan Meyers, Earl Miller, Max Riesenhuber, Dave Perrett, Simon Thorpe, Carlo Tomasi and Antonio Torralba for fruitful discussions and comments on this work.



		d’				Error rates				Hits			
		H	C	M	F	H	C	M	F	H	C	M	F
Experiment 1													
Human	Im.-mask	1.48	1.88	1.52	0.81	0.27	0.22	0.28	0.40	0.68	0.74	0.59	0.35
		0.15	0.20	0.15	0.12	0.02	0.03	0.02	0.02	0.05	0.04	0.05	0.05
Human	ISI 30	2.37	2.52	2.19	1.55	0.16	0.14	0.17	0.27	0.87	0.90	0.80	0.58
		0.19	0.15	0.17	0.14	0.02	0.02	0.02	0.02	0.03	0.02	0.03	0.04
Human	ISI 60	2.69	2.64	2.6	1.84	0.13	0.13	0.13	0.22	0.92	0.92	0.89	0.71
		0.18	0.18	0.15	0.15	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.03
Human	No-mask	3.01	2.82	3.1	2.38	0.10	0.11	0.09	0.15	0.95	0.94	0.94	0.81
		0.21	0.16	0.18	0.16	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.02
Model		2.04	2.48	1.97	1.37	0.18	0.11	0.17	0.26	0.92	0.90	0.79	0.68
		0.07	0.07	0.05	0.05	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01
Experiment 2													
Human	ISI 30	2.20	2.32	2.02	1.45	0.17	0.16	0.20	0.29	0.78	0.82	0.71	0.52
		0.15	0.15	0.14	0.12	0.02	0.02	0.02	0.02	0.04	0.03	0.04	0.04
Model		2.04	2.48	1.97	1.37	0.18	0.11	0.17	0.26	0.92	0.90	0.79	0.68
		0.07	0.07	0.05	0.05	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01
Model V1		1.37	1.78	1.53	0.65	0.26	0.19	0.23	0.38	0.85	0.83	0.78	0.55
		0.04	0.04	0.05	0.04	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Textons		0.84	0.58	0.69	0.35	0.34	0.39	0.37	0.43	0.72	0.62	0.67	0.62
		0.04	0.04	0.04	0.04	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Global Features		1.43	1.73	1.47	0.74	0.25	0.20	0.23	0.36	0.84	0.82	0.75	0.61
		0.05	0.04	0.03	0.05	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table 5.3: Summary of accuracy measures for human observers and the model (continue on next page).

Experiment 3													
Human	0°	2.28	2.39	2.13	1.71	0.15	0.15	0.17	0.25	0.88	0.91	0.83	0.60
		0.22	0.21	0.20	0.15	0.02	0.03	0.02	0.02	0.02	0.02	0.03	0.04
Human	90°	2.15	2.13	1.75	1.13	0.17	0.17	0.22	0.34	0.85	0.86	0.73	0.44
		0.24	0.18	0.19	0.12	0.03	0.02	0.02	0.03	0.30	0.03	0.05	0.05
Human	180°	1.95	2.01	1.96	1.28	0.19	0.18	0.19	0.31	0.82	0.83	0.74	0.51
		0.19	0.19	0.18	0.16	0.02	0.03	0.22	0.02	0.03	0.03	0.04	0.03
Model	0°	2.05	2.35	1.94	1.44	0.20	0.13	0.17	0.24	0.93	0.87	0.80	0.74
		0.11	0.09	0.07	0.05	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02
Model	90°	2.09	2.12	1.34	0.99	0.19	0.16	0.26	0.32	0.92	0.84	0.72	0.62
		0.11	0.10	0.05	0.06	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02
Model	180°	1.99	2.07	1.64	1.25	0.20	0.16	0.21	0.27	0.92	0.85	0.72	0.69
		0.11	0.11	0.08	0.05	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02

Table 5.4: Summary of accuracy measures for human observers and the model (continue from previous page).

Chapter 6

Discussion

A Summary

In this thesis, we have developed a quantitative model of the feedforward pathway of the ventral stream in visual cortex – from cortical area V1 to V2 to V4 to IT and PFC. The 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 and attempts to explain the first few hundred milliseconds of visual processing.

One of the key property of the model 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.

The quantitative nature of the model has allowed us to directly compare properties of the model against experimental observations at three different scales. In Chapter 3, we compared the model against electrophysiological recordings at different levels along the ventral visual stream in the macaque visual system. We have shown that the model is consistent with data from V1, V4 and IT. In Chapter 4 we showed that not only can the model duplicate the tuning properties of neurons in various brain areas when probed with artificial stimuli (like the ones typically used in physiology), but it can also handle the recognition of objects in the real-world, to the extent of competing with the best computer vision systems. In Chapter 5 we compared directly the performance of the model and the performance of human observers in a rapid animal *vs.* non-animal recognition task. Results indicate that the model can predict well not only the level and the pattern of performance of human observers but also the level of difficulty of individual images. Taken together, the evidence presented shows that we may have the skeleton of a successful theory of "immediate recognition" in visual cortex.

B Open Questions

The model certainly does not account for *all* possible visual phenomena and illusions. At best, the theory is just a skeleton still missing many important aspects. Below is an incomplete list of the most obvious open questions.

B.1 The Architecture

How strict is the hierarchy and how precisely does it map into cells of different visual areas? For instance, are cells corresponding to S_2 units in V2 and C_2 units in V4 or are some cells corresponding to S_2 units already in V1? The theory is rather open about these possibilities: the mapping of Fig. 2-1 is just an educated guess. However, because of the increasing arborization of cells and the number of boutons from V1 to PFC [Elston, 2003], the number of subunits to the cells should increase and thus their potential size and complexity. In addition, C units should show more invariance from the bottom to the top of the hierarchy.

What is the nature of the cortical and subcortical connections (both feedforward and feedback) of the main areas of the ventral visual stream that are involved in the model? Such analysis would help improve the architecture of the model by better constraining some of the parameters such as the size of the dictionary of shape-components or the number of inputs to units in different layers. This would also help refine and extend the existing literature on the organization of visual cortex [Felleman and van Essen, 1991]. With the recent apparition of higher resolution tracers (*e.g.,* PHA-L, byocytin, DBA), visualization has greatly improved and it is now possible to go beyond a general layout of interconnected structures and start addressing the finer organization of connections. For instance, recent studies characterized the precise morphology and microstructure of terminal arbors and boutons [see Rockland, 2002]. Fine-scale quantitative characterization of the major brain areas involved in the model is already partly available: this includes the major feedforward routes, *i.e.,* from V1 to V2 [Rockland and Virga, 1990; Girard et al., 2001], V2 to V4 [Gattass et al., 1997], PIT to AIT [Saleem et al., 1993; Steele and Weller, 1995], AIT to STS [Saleem et al., 1996], as well as the feedback connections from V4 and PIT [Rockland et al., 1994; Felleman et al., 1997], and the bypass routes (*i.e.,* V1 to V4 and V2 to PIT [Nakamura et al., 1993]). An analysis should be performed that involve: (1) the likely number of neuron types involved in the first few hundred milliseconds of visual processing and (2) an estimate on the number of afferent inputs for each unit type in the model.

B.2 Learning and Plasticity

What are the precise biophysical mechanisms for the learning rule described in Chapter 2 and how can invariances be learned within the same framework? Possible synaptic mechanisms for learning should be described in biophysical details. As suggested in Chapter 2 there should be at least three different synaptic rules: 1) for learning the TUNING of the units at the S level by detecting correlations between subunits at the same time; 2) for learning the invariance to position and scale at the C level by detecting correlations between subunits across time and 3) for training the task-specific circuits (probably from IT to PFC) in a supervised manner.

Is learning in areas below IT purely unsupervised and developmental-like as assumed in Chapter 2? Or is there task- and/or object-specific learning in adults occurring below IT in V4, V2 or even V1.

B.3 Performance on Natural Images

Have we reached the limit of what a/this feedforward architecture can achieve in terms of performance? In other words, is the somewhat better performance of humans on the animal *vs.* non-animal categorization task over the model for SOAs longer than 80 ms (see Chapter 5) due to feedback effects mediated by the back-projections or is it that the model still need to be improved to attain human performance in the absence of a mask? There could be several directions to follow in order to try to improve the model performance. One possibility would involve experimenting with the size of the dictionary of shape-components (that could be further reduced with feature selection techniques for instance).

Another possibility would involve adding intermediate layers to the existing ones.

Are feedback loops always desirable? Is the performance on a specific task guaranteed to always increase when subjects are given more time? Or are there tasks for which blocking the effect of back-projections with rapid masked visual presentation does increase the level of performance compared to longer presentation times?

C Future Extensions

C.1 The Ventral Pathway

Learning the tuning of the S_1 units. In the present implementation of the model the tuning of the simple cells in V1 is hardwired. It is likely that it could be determined through the same passive learning mechanisms postulated for the $S_2,\ S_{2b}$ and S_3 units (possibly in V4 and PIT), possibly with a slower time scale and constrained to LGN center-surround subunits. We would expect the automatic learning from natural images mostly of oriented receptive fields but also of more complex ones, including end-stopping units (as reported for instance in [DeAngelis et al., 1992] in layer 6 of V1).

Color and stereo mechanisms from V1 to IT should be included. The present implementation deals with gray level images. This fits well with the fact that color information does not seem to impact performance in rapid categorization tasks Delorme et al. [2000]. More complex phenomena involving color such as color constancy and integration of color in visual perception should also be explained. Stereo (along with motion) cues could potentially play a role in unsupervised learning by helping segmenting between the object and the background.

C.2 The Dorsal Pathway

The original model, formerly known as HMAX, was extended to deal with recognition of biological motion and actions [Giese and Poggio, 2003]. Initial work has been done to extend the present theory in the same way. For instance, in [Sigala et al., 2005], we have shown that the addition of a developmental-like learning stage in intermediate stages of a model of the dorsal stream [Giese and Poggio, 2003] also lead to a drastic improvement in terms of recognition performance. Interestingly the corresponding learning rule selected motion features that are critical for human observers to perform the task [see Casile and Giese, 2005]. Such extension is important because the same S_4 units (AIT cells) that we discussed here as supporting recognition of static images are likely to be also part of a network of reciprocal, lateral, local excitatory connections (learned from passive visual experience) and more global inhibition that endows them with sequence selectivity [see Sakai and Miyashita, 1991] and predictivity [Perrett, pers. comm.].

C.3 Including Back-Projections

The most critical extension of the theory has to do with the extensive back-projections in visual cortex which need to be taken into account in any complete theory of visual cortex. In the future, we will have to extend the architecture of the model by including back-projections and assigning meaningful functions to them. Our working hypothesis is that a) difficult recognition tasks, as object categorization in complex natural images, can be done within single "snapshots" (*e.g.,* short visual exposures only require the feedforward architecture of the ventral stream), but b) there are recognition tasks (or levels of performance) that need time: such tasks probably require recursions of predictions and verifications (possibly involving eye or attentional "fixations") and the associated engagement of the back-projection pathways.

Beyond the Feedforward Sweep: Attention, Prediction and Verification

Attentional mechanisms. There is a number of ideas about the role of back-projections. Back-projections may underlie attentional fixations and zooms-in that may be important in improving performance by focusing on specific spots of the image at the relevant scale and position (see [Kanwisher and Wojciulik, 2000] for a review). In this view, one may try to extend the model to perform visual searches and other attentionally demanding processes which are often guided from the top down when a specific task is given [Wolfe et al., 2004] (*i.e.,* account for eye movements and shifts of attention beyond the first 150 milliseconds). Indeed we have recently developed a computational model [Walther et al., 2005], in which V4-like S_2 features are shared between object detection and top-down attention such that by a cascade of feedback connections (from PFC to IT and from IT to V4), top-down processes can re-use these same features to bias attention to locations with higher probability of containing the target object. We showed that the model could perform visual search of faces in natural scenes.

A closely related proposal accounts for receptive field dynamics, such as shrinkage and extension. In this possible extension, the C_2 pooling range (*i.e.*, the number of S_2 units over which the max is taken to compute a C_2 response) is a dynamic variable controlled by feedback connections from IT neurons. This could provide a mechanism for computing the approximate object location from the shape pathway.

Vision with scrutiny. The basic idea – which is not new and more or less accepted in these general terms – is that one key role of back-projections is to select and modulate specific connections in early areas in a top-down fashion – in addition to manage and control learning processes. This highly speculative framework fits best with the point of view described by [Hochstein and Ahissar, 2002]. Its emphasis is thus somewhat different with respect to ideas related to prediction-verification recursions – an approach known in AI as "hypothesis-verification" (see among others, [Hawkins and Blakeslee, 2002; Mumford, 1992; Rao and Ballard, 1999]). Hochstein & Ahissar suggested that explicit vision advances in reverse hierarchical direction, starting with "vision at a glance" (corresponding to our "immediate recognition") at the top of the cortical hierarchy and returning downward as needed in a "vision with scrutiny" mode in which reverse hierarchy routines focus attention to specific, active, low-level units. Of course, there is a large gap between all of these ideas and a quantitative theory of the back-projections such as the one described in this paper for the feedforward path in the ventral stream.

A conceptual framework that tries to make sense of the above set of ideas is the following. A program running in PFC decides, depending on the initial feedforward categorization, the next question to ask in order to resolve ambiguity or improve accuracy. Typically, answering this question involves "zooming in" on a particular subregion of the image at the appropriate level and using appropriate units (for instance at the C_1 level) and calling a specific classifier – out of a repertoire – to provide the answer. This framework involves a flavor of the "20 questions" game and the use of "reverse hierarchy routines" which control access to lower level units.

We have performed a preliminary experiment that suggests that such approach may help improve performance in the animal *vs.* non-animal categorization task described in

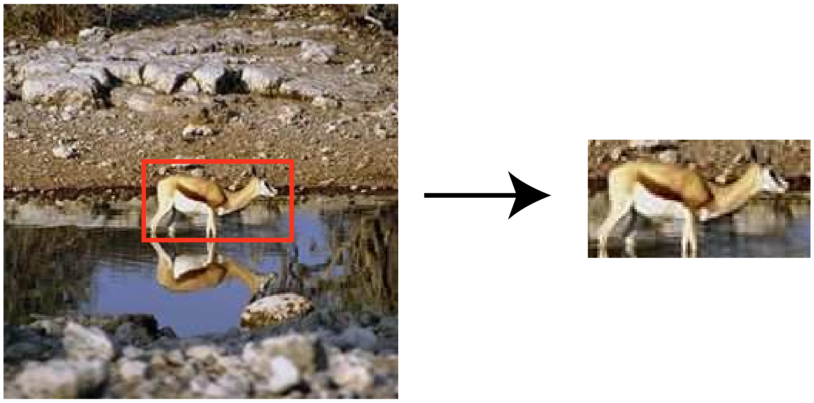


Figure 6-1: An illustration of the "focused" classifier. A first hypothesis about the approximate scale and position of the animal is generated by higher areas during the first feedforward sweep. Back-projections are then used to "zooming in" on a particular subregion of the image at the appropriate level and using appropriate units (for instance at the C_1 level) and calling a specific classifier – out of a repertoire – to provide the answer.

Chapter 5. Fig. 6-1 illustrates the principle: A small window is extracted around the animal and the C_1 units in the corresponding "window of attention" or "'spotlight" [Eriksen and Eriksen, 1974] is passed to a classifier trained on an animal *vs.* non-animal categorization task. Such "focused" classifier does indeed achieve a higher level of performance: On the far-body condition (see Chapter 5) we found an increase in d' from ~ 1.4 to ~ 1.8 (thus reaching the level of human observers in longer SOAs).[1]

Such focused classifier is related to a model for translation (and scale) invariant object recognition put forward several years ago, in the "shifter" circuit by [Anderson and van Essen, 1987] and was later studied by [Olshausen et al., 1993] in a system for attention-based object recognition. A routing circuit, putatively controlled by the pulvinar nucleus in the thalamus, was supposed to re-normalize retinal images to fit into a standard frame of reference which was then used for pattern matching to a store of normalized pictures of objects. Such model could potentially provide an interesting framework to study attentional mechanisms *after* the key, initial feedforward categorization step.

Mental Imagery

Another possible role for back-projections is mental imagery (see [Buckner and Wheeler, 2001] for a review). Preliminary results suggest that it is possible to create mental images within the model under the control of back-projections. In this very speculative proposal, in order to create a mental image of a particular object, *e.g.*, a dog, a vector of neural activity

$\mathbf{X}$ is being synthesized in one of the model layers such that, from the set of all object units[2] in higher brain areas (*e.g.,* a classifier in PFC or a watch unit in the precuneus in the medial parietal area [Fletcher et al., 1995]) only the watch unit will be active. This is illustrated in Fig. 6-2.

Our initial proposal is simple and relies on a stochastic gradient approach. That is, to find a vector of neural activity $\mathbf{X}$ that will cause the right classifier to fire in a higher-order areas, a small perturbation or synaptic noise ϵ is added to the vector of neural activity $\mathbf{X}$. That is the vector of neural activity is modified such that

$$\mathbf{X}' \leftarrow \mathbf{X} + \epsilon.$$

As a result of this change in the vector of neural activity which is propagated through the hierarchy all the way to the top, the activity of the classification units will also change. An increase in the firing of the target classifier will cause the update to be consolidated:

$$\mathbf{X}' = \mathbf{X} + \epsilon$$

else an update in the opposite direction is taken:

$$\mathbf{X}' = \mathbf{X} - \epsilon.$$

Also note that the update only relies on a global feedback signal and could thus be easily controlled through diffuse back-projections.

In what stage of the model should this neural activity $\mathbf{X}$ be synthesized? Intuitively, it makes sense that the higher the stage, the less detailed the synthesized image is. For instance, a mental image produced directly in V4 should contain finer information than a mental image produced in AIT. Conversely the lower the stage is (*i.e.,* the further away from the classification unit), the harder it should be to generate the desired mental image (*i.e.,* for the algorithm to converge) due to the non-linearities at each stage of the model. This is illustrated in Table 6.1. We performed two types of simulations in which $\mathbf{X}$ is generated at the level of the classification units in **cond 1** and at the level of the S_4 units that correspond to the view-tuned units in AIT in **cond 2** (see Fig. 6-2). We ran a small experiment using a subset of 5 of the 101 objects from the *CalTech-101* object dataset.

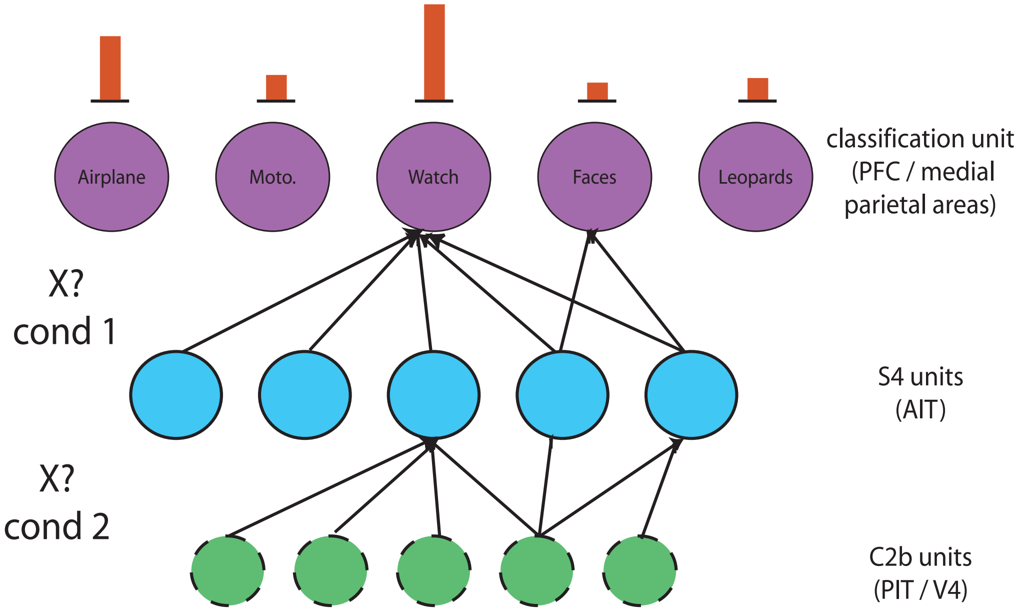


Figure 6-2: A simple model of mental imagery: The assumption is that, for the model to image an object, say a watch the watch classifier and only the watch classifier has to be active. This is done by synthesizing some neural activity in lower model layers (*e.g.,* as an input to the classification units (**cond 1**) or as input to the S_4 units (**cond 2**).

	cond 1		**cond 2**	
	Input to the classification units		Input to the S_4 units (AIT)	
	mean	s.e.m.	mean	s.e.m
Faces	6	1	77	4
Leopards	4	0	105	4
Motorbikes	4	0	36	3
airplanes	6	1	48	4
watch	4	0	58	5

Table 6.1: Mental imagery in the model: Number of feedback loops needed for a mental image to be generated. Comparison between re-activating units at the level of the classification units in higher areas and re-activating units at the level of the S_4 units in AIT. The mean and standard error (s.e.m) were calculated based on 10 runs with random initializations.

In both cases, we found that the algorithm converged, *i.e.,* the right classification unit became active. As expected, the number of feedback loops was much smaller in the **cond 1** where the activity was generated at the level of AIT than in **cond 2** where the neural activity was synthesized lower in PIT/V4 (see Table 6.1). This suggests that creating an image in higher areas should be much faster than in lower areas. We also found that, the algorithm tended to produce a pattern of activity such that the activity of the units that are not selective for the target object is reduced. This seems consistent with a study by O'Craven & Kanwisher [O'Craven et al., 1999] that showed that mental imagery of a particular target object is associated with the subsequent activity of cortical regions selective for this target object. They found that an activation of a cortical region selective for faces (the fusiform face area [Kanwisher et al., 1997] which is likely to correspond to the S_4 units that are tuned to particular face examples in our experiment) when subjects had to create a mental image of a face (compared with imaging places). Conversely they also found a selective activation in the region of the parahippocampal place area [Epstein and Kanwisher, 1998] (that would correspond to S_4 units tuned to places in our experiment) during imagination of places *vs.* faces.

D Predictions

Rapid animal *vs.* non-animal categorization by reading-out from IT: In Chapter 5 we compared the model to human observers on a rapid animal *vs.* non-animal categorization task. We found that for a stimulus asynchrony onset SOA of $50\ ms$, the model could actually predict the level of performance of the human observers very well. To perform the animal classification task the model relies on a linear classifier (probably in PFC) that "looks" at the activity of a few hundred neurons in the S_4 layer corresponding to the view-tuned example-based units from IT. This scheme was motivated by an early proposal by Poggio and Edelman [1990] to explain view-invariant recognition and was closely related to Radial Basis Function (RBF) networks [Poggio and Girosi, 1990]. Interestingly a recent study [Hung et al., 2005] showed that object category can be read-out by a linear classifier from the activity of a few hundreds IT neurons while the monkey is passively viewing images.

A *clear prediction* of the model is that read-out from "IT" for objects in clutter is possible: the simulations on the animal *vs.* non-animal categorization task are with complex natural images with significant clutter. Performance on other databases involving clutter is also very good (see Chapter 4). In particular, we find that the presence of one object can be detected even in the presence of other objects [see Serre et al., 2005a].

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

E Beyond Vision: A Universal Scheme?

Are the model and the principles described in this thesis applicable to other modalities? There are several observations that suggest that there may be a chance that at least some of the ideas described here may generalize to other sensory areas.

For instance, at least within the visual system, it seems that a similar scheme could be generalized to other *substances* (*e.g.,* motion, color, binocular disparity). As suggested by [Adelson and Bergen, 1991] the task of the visual system is to measure the state of the luminous environment or more precisely local changes along various directions in the visual environment. This is illustrated in Fig. 6-3 (reproduced from [Adelson and Bergen, 1991]). For instance, the orientated S_1 units in the model extract useful information about local changes in image intensity along particular $X - Y$ directions. Similar oriented filters in the space and time domain would constitute units that are sensitive for particular directions of motions. Indeed similar units have been used to model motion-sensitive V1 cells in a model of biological motion recognition in the dorsal pathway [Giese and Poggio, 2003]. We have recently extended this model to include an unsupervised developmental-like learning stage similar to the one described in Chapter 2. Not only did the performance of the resulting model increased significantly compared to the original model but it was also showed to learn motion-features that are also used by human observers [Casile and Giese, 2005]. As suggested by [Adelson and Bergen, 1991] and illustrated in Fig. 6-3, it would be very easy to extend the scheme to other substances such as color, disparity, *etc* .

Regarding other sensory modalities, there is a good chance that some of the principles and part of the architecture described in Fig. 2-1 may account also for some of the tuning properties of cells in auditory cortex [T. Ezzat, pers. comm.]. For instance, cortical rewiring experiments have demonstrated that cells in auditory thalamus and cortex from animals in which retinal projections were redirected to the auditory thalamus are visually responsive

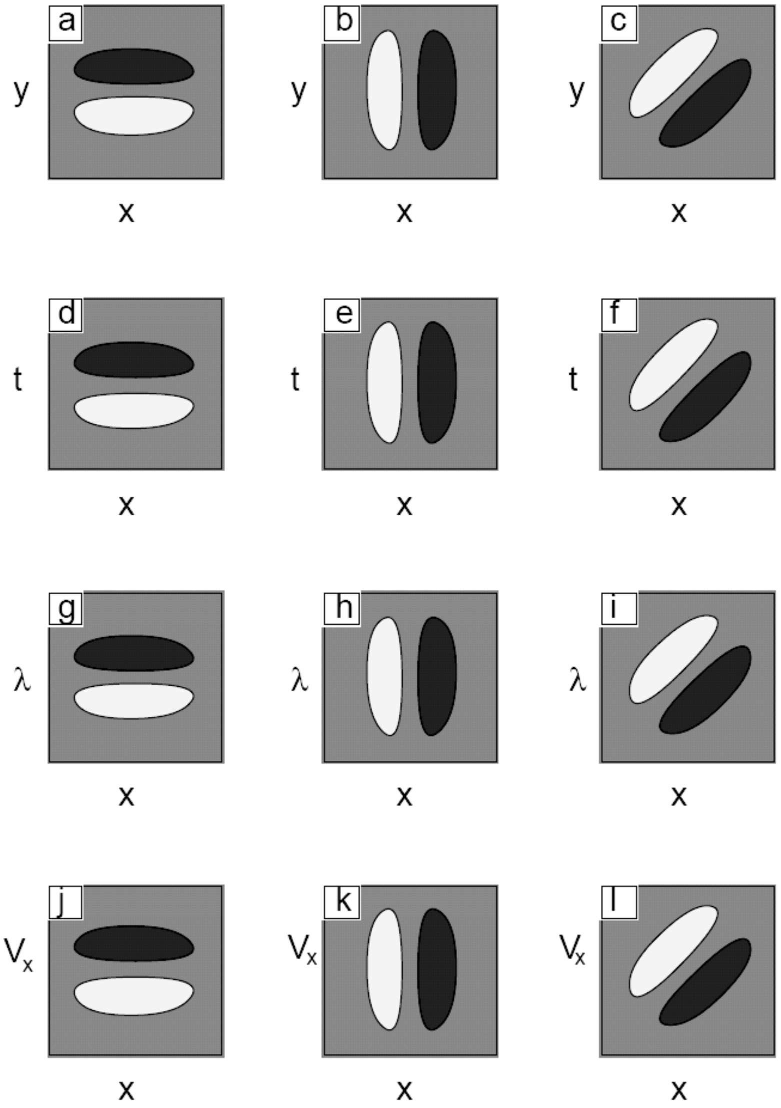


Figure 6-3: The same S_1-type **(a,b,c)** of receptive field structures can produce different measurements when placed along different visual directions (reproduced from [Adelson and Bergen, 1991]). $x,\ y,\ z$ correspond to coordinates in space, λ to the wavelength of the light and $V_x,\ V_y$, and V_z to viewpoint positions.

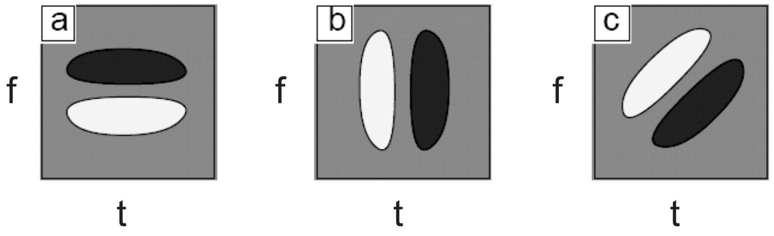


Figure 6-4: S_1 units in the auditory system [T. Ezzat, pers. comm.].

and have receptive field properties that are typical of cells in visual cortex. Additionally it has been shown that this cross-modal projection and its representation in auditory cortex can mediate visual behavior [von Melchner et al., 2000] (see [Newton and Sur, 2004] for a recent review). This may suggest that some of the functional principles may be shared between visual and auditory cortex and that the main differences emerge from differences in the nature of their inputs. Indeed a recent study [Chi et al., 2005] already suggested that oriented receptive field structures of the S_1-type may extend to auditory cortex. This is illustrated in Fig. 6-4.

Finally, in a recent review, [Poggio and Bizzi, 2004] suggested that the motor and visual cortex may share some of the same strategies. In particular, the Gaussian-like TUNING operation Eq. 1.2 may be key in both motor and visual cortex: For instance, some twenty years ago, [Georgopoulos et al., 1982] found neurons that are broadly directionally tuned for arm movements, *i.e.,* their frequency of discharge is a function of the direction of movement, the discharge being strongest along one preferred direction resulting in a directional bell-shaped tuning curve. It has also been reported that in the motor areas of the frontal lobe, neurons with similar preferred direction are interleaved with mini-columns having nearly orthogonal preferred directions [Amirikian and Georgopoulos, 2003] and very similar to the ones described in sensory areas (*e.g.,* visual cortex [see Hubel and Wiesel, 1977], the somato-sensory cortex [Mountcastle, 1957] and the auditory cortex [Merzenich and Brugge, 1973]).

Notes

[1] Another related proposal includes the model of contextual object priming by [Oliva et al., 2003].

[2] In the present version of the model, we have one classification unit for each object class to be recognized. For instance, to perform the experiment on the *CalTech-101* in Chapter 4, the model contained 101 classification units in PFC.

Appendix A

Detailed Model Implementation and Parameters

We here provide a detailed description of the model implementation and of the parameter values. The complete model, corresponding to Fig. 2-1 and described in Chapter 2, which was used in most simulations in Chapter 4 and in Chapter 5), is described below.

The comparison between the model and the benchmark AI systems was performed on a subcomponent of the model which corresponds to the route going from V2 to PIT bypassing V4 (light blue arrows in Fig. 2-1, *i.e.*, layers $S_1 \rightarrow C_1 \rightarrow S_{2b} \rightarrow C_{2b} \rightarrow$ *PFC classifier*, see [Serre and Riesenhuber, 2004; Serre et al., 2005b, 2006b]). This was shown to give a good compromise between speed and accuracy. Matlab code for this model subcomponent can be found at `http://cbcl.mit.edu/software-datasets/standardmodel/index.html`.

A Model Architecture and Implementation

There are two types of functional layers in the model: the S layers which are composed of *simple* units are interleaved with C layers which are composed of *complex* units.

Simple units in the S_k layer pool over afferent units from a topologically related local neighborhood in the previous C_{k-1} layer with different selectivities. As a result, the complexity of the preferred stimulus of units increases from layer C_{k-1} to S_k. The pooling operation at the S level is a Gaussian-like tuning function. That is, the response y of a simple unit, receiving the pattern of synaptic inputs $\left(x_1,\ \ldots,\ x_{n_{S_k}}\right)$ from the previous layer is given by:

$$y = \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n_{S_k}}(w_j - x_j)^2\right), \tag{A.1}$$

where σ defines the sharpness of the TUNING around the preferred stimulus of the unit corresponding to the weight vector $\mathbf{w} = (w_1,\ \ldots\ w_{n_{S_k}})$. That is, the response of the unit is maximal ($y = 1$) when the current pattern of input $\mathbf{x}$ matches exactly the synaptic weight vector $\mathbf{w}$ and decreases with a bell-shaped tuning profile as the pattern of input becomes more dissimilar.[1]

Complex units in the C_k layer pool over afferent units from the previous S_k layer with the same selectivity but at slightly different positions and scales to increase the tolerance to 2D transformations from layer S_k to C_k. The pooling operation at the complex C level is a MAX operation. That is, the response y of a complex unit corresponds to the response of the strongest of its afferents $\left(x_1,\ \ldots,\ x_{n_{C_k}}\right)$ from the previous S_k layer. An idealized mathematical description of the complex unit operation is given by:

$$y = \max_{j=1\ldots\ n_{C_k}} x_j. \tag{A.2}$$

A complete description of the two operations, a summary of the evidence as well as plausible biophysical circuits to implement them can be found in [Serre et al., 2005a, Section 5, pp. 53-59].

Functional organization: Layers in the model are organized in *feature maps* which may be thought of as *columns* or *clusters* of units with the *same selectivity* (or preferred stimulus) but with receptive fields at slightly different scales and positions (see Fig. 2-6). Within one feature map all units share the same selectivity, *i.e.,* synaptic weight vector $\mathbf{w}$ which is learned from natural images (see Chapter 2).

There are several parameters governing the organization of individual layers: K_X is the number of feature maps in layer X. Units in layer X receive their inputs from a topologically related $\Delta N_X \times \Delta N_X \times \Delta S_X$, grid of possible afferent units from the previous layer where ΔN_X defines a range of positions and ΔS_X a range of scales.

Simple units pool over afferent units at the same scale, *i.e.*, ΔS_{S_k} contains only a single scale element. Also note that in the current model implementation, while complex units pool over all possible afferents such that each unit in layer C_k receives $n_{C_k} = \Delta N^S_{C_k} \times \Delta N^S_{C_k} \times \Delta S_{C_k}$, simple units receive only a subset of the possible afferent units (selected at random) such that $n_{S_k} < \Delta N_{S_k} \times \Delta N_{S_k}$ (see Table A.1 for parameter values).

Finally, there is a downsampling stage from S_k to C_k stage. While S units are computed at all possible locations, C units are only computed every ϵ_{C_k} possible locations. Note that there is a high degree of overlap between units in all stages (to guarantee good invariance to translation). The number of feature maps is conserved from S_k to C_k stage, *i.e.*, $K_{S_k} = K_{C_k}$. The value of all parameters is summarized in Table A.1.

S_1 **and** C_1 **stages:** The input to the model is a still[2] gray-value image ($256 \times 256 \sim 7^o \times 7^o$ of visual angle) which is first analyzed by a multi-dimensional array of simple S_1 units which correspond to the classical V1 simple cells of Hubel & Wiesel. 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). Fig. 2-2 shows the different weight vectors corresponding to the different types of S_1 units (only one phase shown). Mathematically the weight vector $\mathbf{w}$ of the S_1 units take the form of a Gabor function [Gabor, 1946], which have been shown to provide a good model of simple cell receptive fields [Jones and Palmer, 1987] and can be described by the following equation:

$$F(u_1, u_2) = \exp\left(-\frac{(\hat{u_1}^2 + \gamma^2 \hat{u_2}^2)}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda}\hat{u_1}\right), \quad \text{s.t.} \tag{A.3}$$

$$\hat{u_1} = u_1 \cos\theta + u_2 \sin\theta \quad \text{and} \tag{A.4}$$

$$\hat{u_2} = -u_1 \sin\theta + u_2 \cos\theta, \tag{A.5}$$

The five parameters, *i.e.*, orientation θ, aspect ratio γ, effective width σ, phase ϕ and wavelength λ determine the properties of the spatial receptive field of the units. The tuning of simple cells in cortex varies substantially along these dimensions. We consider four orientations ($\theta = 0°$, 45° , 90° , and 135°). This is an over-simplification but this was previously shown to be sufficient to provide rotation and size invariance at the S_4 level in good agreement with recordings from AIT Riesenhuber and Poggio [1999a]. ϕ was set to 0° while different phases are crudely approximated by centering receptive fields at all locations. In order to obtain receptive field sizes consistent with values reported for parafoveal simple cells Schiller et al. [1976e], we considered 17 filters sizes from 7×7 (0.2° visual angle) to 39×39 (1.1° visual angle) obtained by steps of two pixels.

When fixing the values of the remaining 3 parameters (γ, λ and σ), we tried to account for general cortical cell properties, that is:

1. The peak frequency selectivity of cortical cells tends to be negatively correlated with the sizes of the receptive fields Schiller et al. [1976d]

2. The spatial frequency selectivity bandwidth of cortical cells tends to be positively correlated with the sizes of the receptive fieldsSchiller et al. [1976d]

3. The orientation bandwidth of cortical cells tends to be positively correlated with the sizes of the receptive fields Schiller et al. [1976c].

We empirically found that one way to account for all three properties is to include fewer cycles in the receptive fields of the units as their sizes (*R*F size) increase. We found that the two following (ad hoc) formulas gave good agreement with the tuning properties of cortical cells:

$$\sigma = 0.0036 * \mathit{RF\ size}^2 + 0.35 * \mathit{RF\ size} + 0.18 \qquad \text{(A.6)}$$

$$\lambda = \frac{\sigma}{0.8} \qquad \text{(A.7)}$$

For all cells with a given set of parameters (λ_0, σ_0) to share similar tuning properties at all orientations, we applied a circular mask to the receptive field of the S_1 units. Cropping Gabor filters to a smaller size than their effective length and width, we found that the aspect ratio γ had only a limited effect on the cells tuning properties and was fixed to 0.3 for all filters.

The next C_1 level corresponds to striate complex cells [Hubel and Wiesel, 1959]. Each of the complex C_1 units receives the outputs of a group of simple S_1 units with the same preferred orientation (and two opposite phases) but at slightly different positions and sizes (or peak frequencies). The result of the pooling over positions is that C_1 units become insensitive to the location of the stimulus within their receptive fields, which is a hallmark of the complex cells [Hubel and Wiesel, 1959]. As a result, the size of the receptive fields increase from the S_1 to the C_1 stage (from $0.2^o - 1.0^o$ to $0.4^o - 2.0^o$). Similarly the effect of the pooling over scales 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].

The parameters of the Gabor filters (see Eq. A.3) were adjusted so that the tuning properties of the corresponding S_1 units match closely those of V1 parafoveal simple cells [Serre et al., 2004b]. Similarly the pooling parameters at the next stage were adjusted so that the tuning and invariance properties of the corresponding C_1 units match closely those of V1 parafoveal complex cells.[3] The complete parameter set used to generate the population of S_1 units is given in Table A.1.

S_2 and C_2 stages: At the S_2 level, units pool the activities of $n_{S_2} = 10$ retinotopically organized complex C_1 units at different preferred orientations over a $\Delta N_{S_2} \times \Delta N_{S_2} = 3 \times 3$ neighborhood of C_1 units via a TUNING operation. As a result, the complexity of the preferred stimuli is increased: At the C_1 level units are selective for single bars at a particular orientation, whereas at the S_2 level, units become selective to more complex patterns – such as the combination of oriented bars to form contours or boundary-conformations. Receptive field sizes at the S_2 level range between $0.6^o - 2.4^o$.

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 field. Receptive field sizes at the C_2 level range between $1.1^o - 3.0^o$.

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], see Chapter 3.

S_3 and C_3 stages: Beyond the S_2 and C_2 stages 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), where the response of $n_{S_3} = 100$ C_2 units 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 Table A.1) 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 [Serre et al., 2004b, 2005a]). 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 is about the size of the stimulus (from $4^o \times 4^o$ to $7^o \times 7^o$).

S_{2b} and C_{2b} stages: They may correspond to the bypass routes that have been found in visual cortex, *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]. 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 their receptive field is larger (2 to 3 times larger) than the receptive fields of the S_2 units. Importantly, the number of afferents to the S_{2b} units is also larger ($n_{S_{2b}} = 100$ *vs.* $n_{S_2} = 10$), 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} also results at the C_{2b} level in units that are more selective than other units at a similar level along the hierarchy (C_3 units), and at the same time exhibit a smaller range of invariance to positions and scales. We found that the tuning of the C_{2b} units agree with the read out data from IT [Hung et al., 2005] (see [Serre et al., 2005a]).

Biophysical implementations of the key computations: The model implementation used here is agnostic about the implementations of the Gaussian-like tuning and the max-like operations as well as about the biophysical mechanisms of unsupervised and supervised learning. For the two key computations we used the idealized operations described in Eq. A.2 and Eq. A.1. There are plausible local circuits [Serre et al., 2005a] implementing the two key operations within the time constraints of the experimental data [Perrett et al.,

	S_1 parameters							
RF size (pixels)	7 & 9	11 & 13	15 & 17	19 & 21	23 & 25	27 & 29	31 & 33	35 & 37 & 39
σ	2.8 & 3.6	4.5 & 5.4	6.3 & 7.3	8.2 & 9.2	10.2 & 11.3	12.3 & 13.4	14.6 & 15.8	17.0 & 18.2 & 19.5
λ	3.5 & 4.6	5.6 & 6.8	7.9 & 9.1	10.3 & 11.5	12.7 & 14.1	15.4 & 16.8	18.2 & 19.7	21.2 & 22.8 & 24.4
θ	$0^0; 45^0; 90^0; 180^0$							
num. S_1-types K_{S_1}	4							
	C_1 parameters							
Bands ΔS_{C_1}	1	2	3	4	5	6	7	8
grid size $\Delta N^S_{C_1}$	8	10	12	14	16	18	20	22
sampling ϵ_{C_1}	3	5	7	8	10	12	13	15
num. C_1-types K_{C_1}	$= K_{S_1} = 4$							
	S_2 parameters							
grid size ΔN_{S_2}	3×3 ($\times 4$ orientations)							
num. afferents n_{S_2}	10							
num. S_2-types K_{S_2}	≈ 2000							
	C_2 parameters							
Bands ΔS_{C_2}	1 & 2		3 & 4		5 & 6		7 & 8	
grid size $\Delta N^S_{C_2}$	8		12		16		20	
sampling ϵ_{C_2}	3		7		10		13	
num. C_2-types K_{C_2}	$= K_{S_2} \approx 2000$							
	S_3 parameters							
grid size ΔN_{S_3}	3×3 ($\times K_{S_2}$)							
num. afferents n_{S_3}	100							
num. S_3-types K_{S_3}	≈ 2000							
	C_3 parameters							
Bands ΔS_{C_3}	1 & 2 & 3 & 4 & 5 & 6 & 7 & 8							
grid size $\Delta N^S_{C_3}$	40							
num. C_3-types K_{C_3}	$= K_{S_3} \approx 2000$							
	S_{2b} parameters							
grid size $\Delta N_{S_{2b}}$	6×6; 9×9; 12×12; 15×15 ($\times 4$ orientations)							
num. afferents $n_{S_{2b}}$	100							
num. S_{2b}-types $K_{S_{2b}}$	$\approx$ 500 for each size $\approx$ 2000 total							
	C_{2b} parameters							
Bands $\Delta S_{C_{2b}}$	1 & 2 & 3 & 4 & 5 & 6 & 7 & 8							
grid size $\Delta N^S_{C_{2b}}$	40							
num. C_{2b}-types $K_{C_{2b}}$	$= K_{S_{2b}} \approx$ 500 for each size$\approx$ 2000 total							

Table A.1: Summary of all the model parameters.

1992; Hung et al., 2005] based on small local population of spiking neurons firing probabilistically in proportion to the underlying analog value [Smith and Lewicki, 2006] and on shunting inhibition [Grossberg, 1973]. Other possibilities may involve spike timing in individual neurons (see [VanRullen et al., 2005] for a recent review).

B Major Extensions from the Original HMAX model

The architecture sketched in Fig. 2-1 has evolved – as originally planned and from the interaction with experimental labs – from the original model by [Riesenhuber and Poggio, 1999a]. In particular, new layers have been added (now accounting for V4 and PIT in separate layers for instance) to improve the mapping between the functional primitives of the theory and the structural primitives of the ventral stream in the primate visual system. Below is a list of major changes and differences between this new model implementation and the original one:

1. S_1 **and** C_1 **layers:** In [Serre and Riesenhuber, 2004] we found that the S_1 and C_1 units in the original model were too broadly tuned in terms of orientation and spatial frequency and proposed a new set of units that better capture the tuning properties of V1 cortical cells. In particular at the S_1 level, we replaced Gaussian derivatives with Gabor filters which we found more suited to fit V1 data. We also modified the receptive field sizes and tuning properties of both S_1 and C_1 units.

2. S_2 **layer:** The tuning of the S_2 units is now learned from natural images (see Chapter 2). S_2 units are more elaborate than the S_2 units in the original HMAX (simple 2×2 combinations of orientations). *The introduction of learning, we believe, has been a key factor for the model to achieve a high level of performance on the recognition of complex images* (see [Serre et al., 2002, 2005b, 2006b] and Chapter 4).

3. C_2 **layer:** The receptive field size of the C_2 units, as well as the range of invariances to scale and position is now reduced such that C_2 units better fit V4 data. See Chapter 3 for details.

4. S_3 **and** C_3 **layers:** These two layers were added only recently and constitute the topmost layers of the model along with the S_2b and C_2b units (see Chapter 2 and above). The tuning of the S_3 units is also learned from natural images.

5. S_{2b} **and** C_{2b} **layers:** We added these two layers to account for the bypass route (that projects directly from V1/V2 to PIT, thus bypassing V4 [see Nakamura et al., 1993]). Interestingly these bypass routes have been shown to provide an excellent compromise (when used alone) between speed and accuracy in computer vision applications (see [Serre et al., 2005b, 2006b]).

Notes

[1] When Eq. A.1 is approximated by a normalized dot-product followed by a sigmoid, such that:

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

the weight vector $\mathbf{w}$ corresponds to the strength of the synaptic inputs to the Gaussian-tuned unit.

[2] The present version of the model deals with one single image at a time as it does not incorporate mechanisms for motion and the recognition of sequences. A natural extension to include time may start with a version of the original HMAX model that had the capability of recognizing image sequences [Giese and Poggio, 2003].

[3] Unlike in [Riesenhuber and Poggio, 1999a], all the V1 parameters here are derived exclusively from available V1 data and do not depend as they did in part in [Riesenhuber and Poggio, 1999a] from the requirement of fitting the benchmark paperclip recognition experiments [Logothetis et al., 1995]. Thus the fitting of these paperclip data by the model is even more remarkable than in [Riesenhuber and Poggio, 1999a].

[4] In the model, both the supervised and unsupervised learning stages are relatively fast. Yet at run-time, it takes about one minute to classify a single image. A speed up by a factor of 10 is feasible.

Appendix B

Additional Comparisons with Computer Vision Systems

As pointed out in Chapter 2, we only used a subpart of the model for this comparison (*i.e.,* the bypass route depicted on Fig. 2-1), which contains the path running from $S_1 \rightarrow C_1 \rightarrow S_{2b} \rightarrow C_{2b}$. The C_{2b} unit responses were then passed to a linear classifier (boosting or SVM). This gave a good compromise between speed and accuracy in this application-oriented setting with large real-world image databases.

The details of the model implementation are given in Appendix A. We show two applications of the model to computer vision: Semi-supervised object recognition in clutter, for which training is performed on unsegmented images (*i.e.,* the object is present in clutter) and a scene-understanding system. Part of this work appeared in various forms in [Serre et al., 2004b, 2005b, 2006b].

A Image Datasets

We tested the model on various object categorization tasks for comparison with benchmark computer vision systems. All datasets used contain images for which the target object is present or absent.

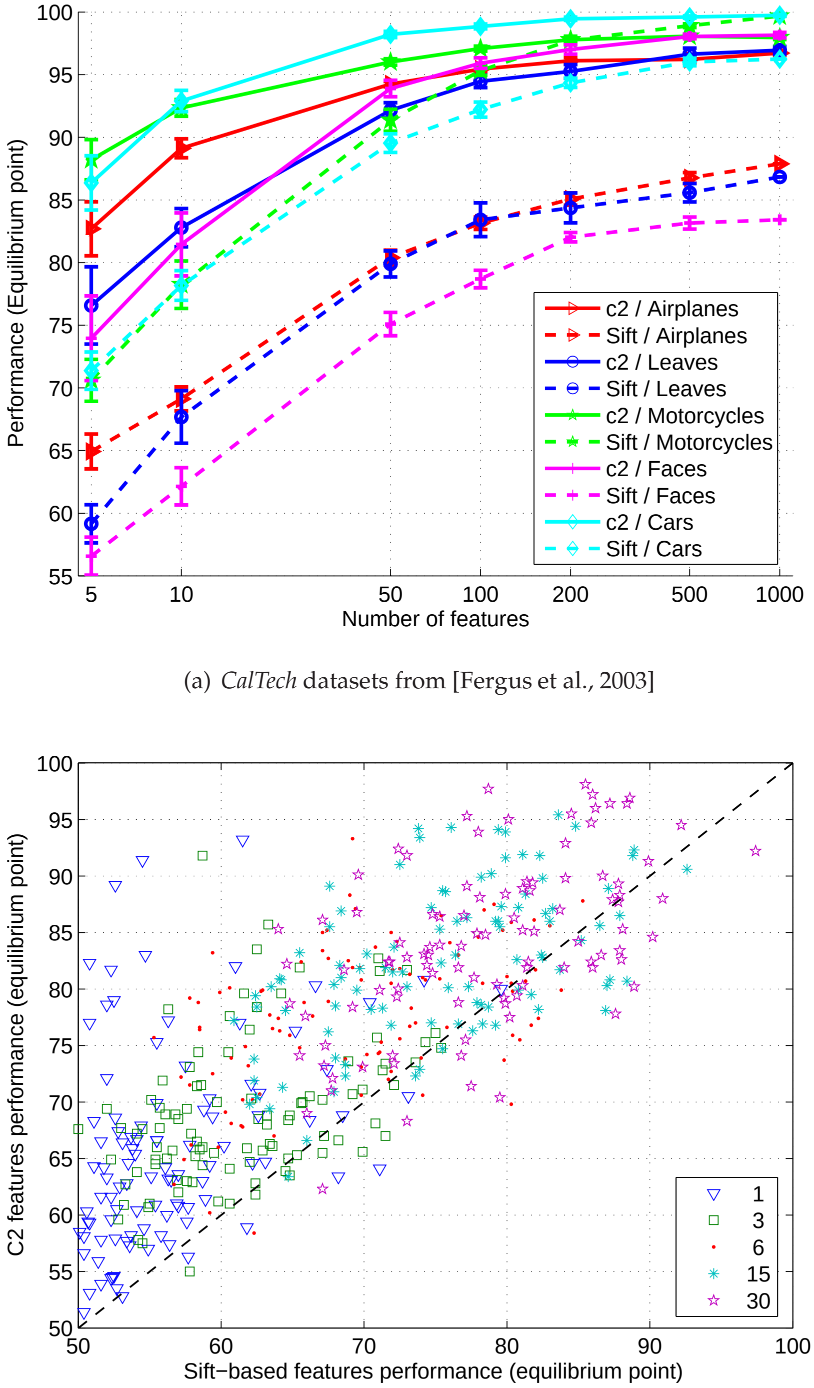
CalTech-5: We consider five databases from the CalTech vision group[1], *i.e.,* frontal-face, motorcycle, rear-car and airplane datasets from [Fergus et al., 2003], as well as the leaf dataset from [Weber et al., 2000b] (see Fig. 4-6 for examples). On these datasets, we used the same fixed splits as in the corresponding studies whenever applicable and otherwise generated random splits. All images were rescaled to be 140 pixels in height (width was rescaled accordingly so that the image aspect ratio was preserved) and converted to grayscale.

CalTech-101: The *CalTech-101* contains 101 object classes plus a background class (see [Fei-Fei et al., 2004] for details, Fig. 4-1 and Fig. 4-3). All results reported were generated with 10 random splits. For training, we used 50 negative examples and a variable number of positive training examples (1, 3, 15, 30 and 40). For testing, in the binary classification experiments we selected 50 negative examples and as many as 50 positive examples from the remaining images. In the multi-class experiment, we used as many as 50 examples per class. All images were rescaled to be 140 pixels in height (width was rescaled accordingly so that the image aspect ratio was preserved) and converted to grayscale.

MIT-CBCL: This includes a near-frontal ($\pm 30^\circ$) face dataset [Heisele et al., 2002] and a multi-view car dataset from [Leung, 2004] (see Fig. 4-7). The face dataset contains about 6,900 positive and 13,700 negative images for training and 427 positive and 5,000 negative images for testing. The car dataset contains 4,000 positive and 1,600 negative training examples and 1,700 test examples (both positive and negative). Although the *benchmark* algorithms were trained on the full sets and the results reported accordingly, our system only used a subset of the training sets (500 examples of each class only).

B Results

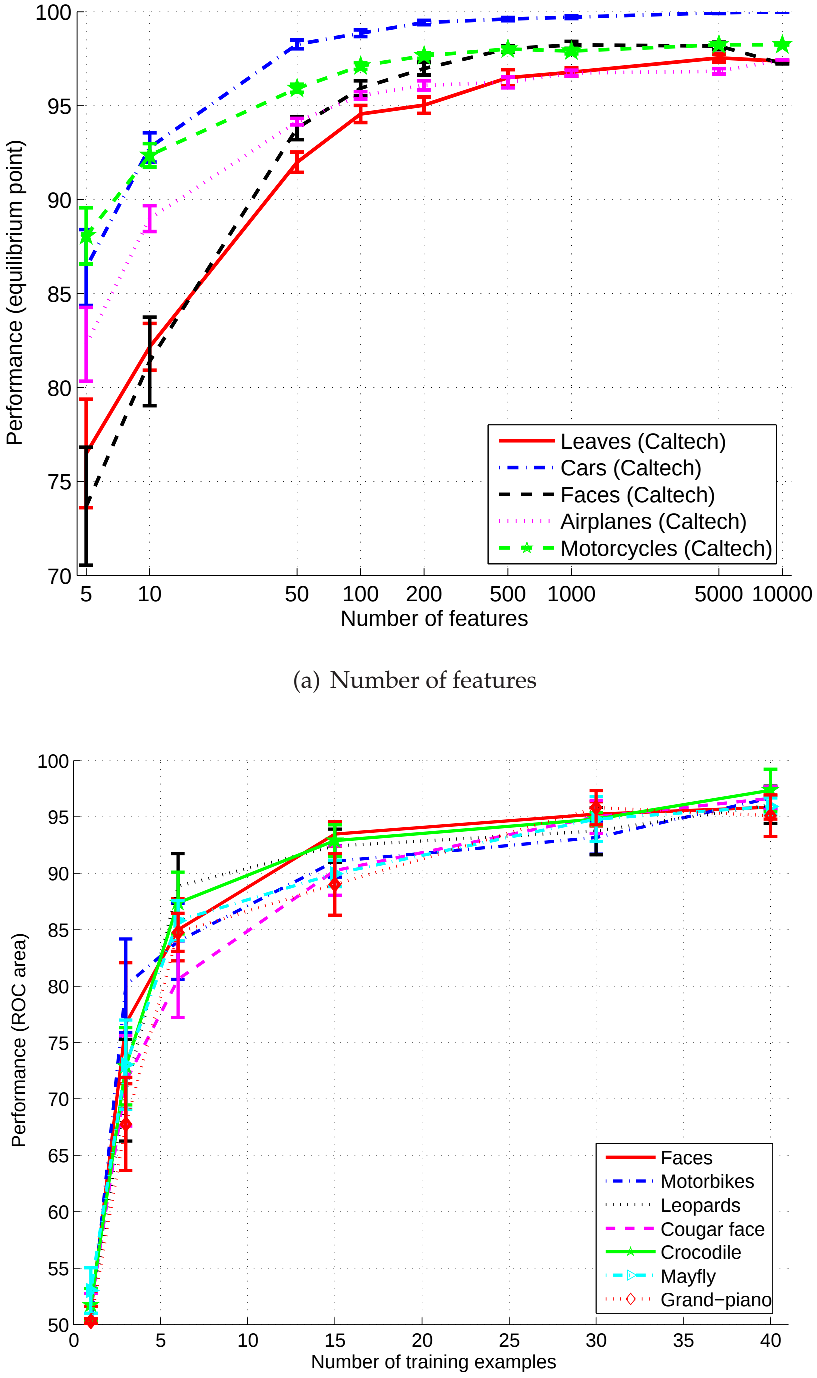
Comparison with SIFT features: We also compared the C_{2b} features to a system based on Lowe's SIFT features [Lowe, 1999]. To perform this comparison at the feature level and ensure a fair comparison between the two systems, we neglected all position information recovered by Lowe's algorithm. It was recently suggested in [Lazebnik et al., 2005] that structural information does not seem to help improve recognition performance. We selected $1,000$ random reference key-points from the training set. Given a new image, we measured the minimum distance between all its key-points and the $1,000$ reference key-points, thus obtaining a feature vector of size $1,000$.[2]



(a) *CalTech* datasets from [Fergus et al., 2003]

(b) *CalTech-101* dataset from [Fei-Fei et al., 2004]

Figure B-1: Comparison between a linear classifier that uses the response of the C_{2b} model units as an input *vs.* the SIFT features [Lowe, 2004]. **(a)** Comparison on the *CalTech-5* datasets [Fergus et al., 2003] for different number of features used. **(b)** Comparison on the *CalTech-101* object database for different numbers of examples available for training.



(a) Number of features

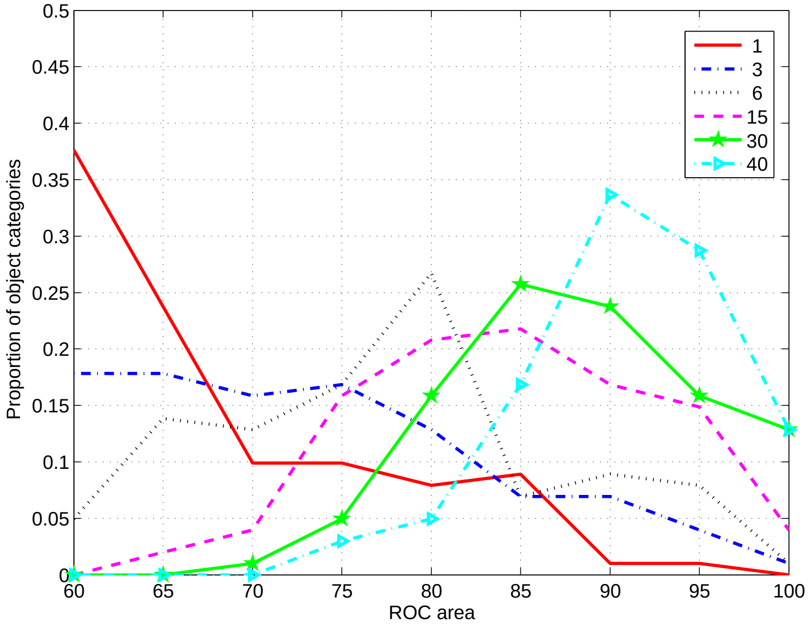
(b) Number of training examples

Figure B-2: Performance *vs.* size of the dictionary of C_{2b} units on the *CalTech-5* datasets [Fergus et al., 2003] **(a)** and on the number of positive examples available for training on sample object category from the *CalTech-101* object dataset **(b)**.

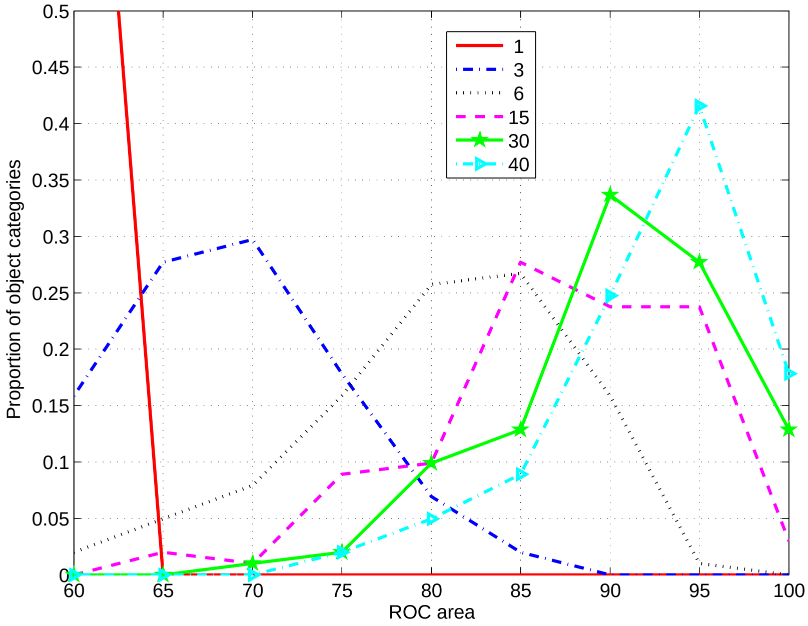
Fig. B-1 shows a comparison between the performance of the SIFT and the C_{2b} features (both with gentleBoost but similar results were obtained with a linear SVM). Fig. B-1(a) shows a comparison on the *CalTech-5* for different number of features and Fig. B-1(b) on the *CalTech-101* database for different number of training examples. In both cases the C_{2b} features outperform the SIFT features significantly. SIFT features excel in the re-detection of a transformed version of a previously seen example but may lack selectivity for a more general categorization task at the basic level.

Number of features and training examples: To investigate the contribution of the number of features on performance, we first created a set of $10,000$ C_{2b} features and then randomly selected subsets of various sizes. The results reported are averaged over 10 independent runs. As Fig. B-2(a) shows, while the performance of the system can be improved with more features (*e.g.,* the whole set of $10,000$ features), reasonable performance can already be obtained with $50 - 100$ features. Interestingly, the number of features needed to reach the plateau (about $1,000 - 5,000$ features) is much larger than the number used by current systems (on the order of 10-100 for [Ullman et al., 2002; Heisele et al., 2002; Torralba et al., 2004] and 4-8 for constellation approaches [Weber et al., 2000b; Fergus et al., 2003; Fei-Fei et al., 2004]). This may come from the fact that we only sample the space of features and do not perform any clustering step like other approaches (including an earlier version of this system [Serre et al., 2002]), we found it to be sensitive to the choice of parameters and initializations, leading to poorer results.

We also studied the influence of the number of training examples on the performance of the system on the *CalTech-101*. For each object category, we generated different positive training sets of size 1, 3, 6, 15 and 30 as in [Fei-Fei et al., 2004]. As shown in Fig. B-2(b) the system achieves error rates comparable to [Fei-Fei et al., 2004] on few training examples (less than 15) but its performance still improves with more examples (where the system by Fei-Fei *et al.* seems to be reaching a plateau). Results with an SVM (not shown) are similar, although the performance tended to be higher on very few training examples (as SVM seems to avoid overfitting even for one example). However, since SVM does not *select* the relevant features, its performance tends to be lower than gentleBoost as the number of training examples increases.



(a) Linear SVM classifier



(b) GentleBoost classifier

Figure B-3: Overall performance on the *CalTech-101* for two types of linear classifiers: **(a)** SVM and **(b)** gentleBoost. Each plot is an histogram of the mean performance of the system across all the 101 different object categories and for different numbers of positive training examples.

Fig. B-3 shows the performance of the gentleBoost and SVM classifiers used with the C_{2b} features on all categories and for various number of training examples (each result is an average of 10 different random splits). Each plot is a single histogram of all 101 scores, obtained using a fixed number of training examples, *e.g.*, with 40 examples, the gentleBoost-based system gets around 95% ROC area for 42% of the object categories.

Multiclass results on the CalTech-101: Finally, we report results on multi-class classification on the *CalTech-101*. To conduct this experiment we use a small dictionary of just $1,000$ features. The classifier is a multi-class linear SVM that applied the all-pairs method, and is trained on 102 labels (101 categories plus the background category). We split each category into a training set of size 15 and a test set containing up to 50 images. Performance is then averaged across all categories. The performance of the system reaches above 44% correct classification rate (chance $< 1\%$) when using 15 training examples per class averaged over 10 repetitions (s.t.d of 1.14%). Using only 5 training images per class, the performance degrades to $\sim 30\%$.

By enlarging the dictionary of shape-components and computing additional *gestalt*-like features (*e.g.*, good-continuity detectors, circularity detectors and symmetry detectors) within the same framework, Wolf & Bileschi obtained $\approx 51.2\% \pm 1.2\%$ correct [Wolf et al., 2006; Bileschi and Wolf, 2006]. Extending our approach, Mutch & Lowe reported 56% correct by applying a feature selection method on the set of C_{2b} features [Mutch and Lowe, 2006]. Some of the best systems include the system by [Holub et al., 2005b] ($\approx 44\%$ correct) and the system by [Berg et al., 2005] (45% correct).

Scene Understanding with the Model: Recently Bileschi & Wolf applied the model to the recognition of complex visual scenes. Outdoor images of cities and suburbs were selected as an appropriate domain for the scene-understanding system. A database of nearly $10,000$ high-resolution images has been collected, more than $3,000$ of which have been hand labeled for 9 object categories. Sample images, their hand labellings, and some empirical results are illustrated in Fig B-4.[3]

The system is composed of two parts: One subcomponent deals *shape-based objects*, the other with *texture-based objects*. Shape-based objects are those objects for which there exists a strong part-to-part correspondence between examples, including pedestrians, cars,

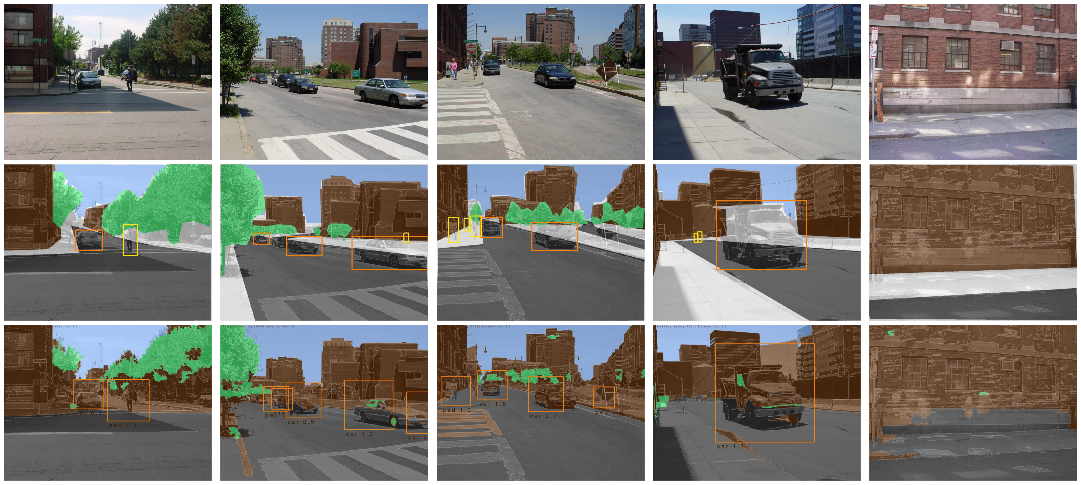


Figure B-4: Sample results by the StreetScene recognition system by Bileschi & Wolf. Top Row: Sample StreetScenes examples. Middle Row: True hand-labeling; color overlay indicates texture-based objects and bounding rectangles indicate shape-based objects. Note that pixels may have multiple labels due to overlapping objects. Bottom Row: Results obtained with the system.

and bicycles. In order to detect shape-based objects, a standard windowing technique is used. This contrasts with the approach described in Chapter 2, wherein isolated objects in clutter are detected using scale- and translation-invariant features, rather than testing for object presence at each position and scale independently. The windowing approach used in this computer vision system may be thought of as the skeleton of an attentional circuit.[4] In conjunction with this windowing approach, we use the C_1 units. Since the window crops away much of the clutter, leaving the potential object nearly centered, the additional invariance from higher model stages is not necessary. It is important to note that the good performance of the C_1 features is dependent upon training data with accurate descriptions of the position and scale of the target object.

Texture-based objects, on the other hand, are those objects for which, unlike shape-based objects, there is no obvious visible inter-object part-wise correspondence. These objects are better described by their texture rather than the geometric structure of reliably detectable parts. For the StreetScenes database these currently include buildings, roads, trees, and skies. The detection of the texture-based objects begins with the segmentation of the input-image*Edison* software [Christoudias et al., 2002]. Segments are assigned labels by calculating the C_{2b} responses within each segment, and inputting this vector into a classifier. One classifier is trained for each object-type using examples from the training database.

C Summary of All Comparisons with Computer Vision Systems

Table B.1 summarizes several comparisons between the model and other state-of-the-art computer vision systems. For this comparison, an earlier (simpler) implementation of the model [Serre et al., 2005b], which corresponds to the *bypass* route projecting from $S_1 \rightarrow C_1 \rightarrow S_{2b} \rightarrow C_{2b}$, was used. The performance of the full architecture which includes a richer dictionary of shape components, tends to be significantly higher than the performance of this simpler (incomplete) implementation. Therefore the results reported here constitute a lower bound on the system performance. These comparisons are based on three studies[5]:

- In [Serre et al., 2005b] we compared the model to the constellation models [Weber et al., 2000b; Fergus et al., 2003] on five standard publicly available datasets from the Caltech vision group: Leave (Lea), Car, Face (Fac), Airplane (Air) and Motorcycle (Mot) as well as two other component-based systems [Heisele et al., 2002; Leung, 2004] on the MIT-CBCL Face (Fac) and Car datasets.

- [Chikkerur and Wolf, 2006] re-implemented the fragment-based system by Ullman and colleagues [Ullman et al., 2002; Epshtein and Ullman, 2005] for comparison with the model on five publicly available datasets: the Leave, Face and Motorcycle datasets from CalTech and the Cow and Face dataset from the Weizmann Institute.

- [Bileschi and Wolf, 2005] re-implemented several systems for comparison with the model on the MIT-CBCL Street Scene dataset. They re-implemented two object recognition systems [Torralba et al., 2004; Leibe et al., 2004] for comparison on the "shape-based" object categories, *i.e.,* Bike (Bik), Pedestrian (Ped), and Car as well as two texture recognition systems [Renninger and Malik, 2004; Carson et al., 1999] for comparison on the "texture-based" object categories, *i.e.,* Building (Bui), Tree (Tre), Road (Roa) and Sky.

In Table B.1, blue indicates that the corresponding study [Serre et al., 2005b] relied on published results of the benchmark systems on standard datasets. Yellow indicates that the results for the benchmark systems were based on re-implementations by the authors of the studies [Bileschi and Wolf, 2005; Chikkerur and Wolf, 2006]. In the study by [Bileschi and Wolf, 2005] the two numbers for the model on Bike, Pedestrian and Car correspond to the performance of the model C_{2b} and C_1 units respectively.[6]

Notes

[1] The *CalTech-5* databases are publicly available at:
`http://www.robots.ox.ac.uk/~vgg/data3.html`.

[2] Lowe recommends using the ratio of the distances between the nearest and the second closest key-point as a similarity measure. We found instead that the minimum distance leads to better performance than the ratio.

[3] This database will soon be available online at:
`http://cbcl.mit.edu/software-datasets`, [see Serre et al., 2006b] for details.

[4] While the purely feedforward approach is appropriate for fast decisions of object presence or absence, it would be impractical for this scene-understanding application as the locations of individual objects would be lost. The windowing approach, however, requires the manual segmentation and normalization of the training set of examples.

[5] Mutch & Lowe also reports favorable comparison in their implementation of the model [Mutch and Lowe, 2006].

[6] On these datasets, images are aligned and normalized, and the amount of clutter is minimal. For such tasks, for which there is no variation of the object in shift and scale, lower stages of the model (*e.g.*, C_1 stage) tend to perform better than higher stages (*e.g.*, C_{2b}).

		Weizmann		CalTech					MIT-CBCL	
		Fac	Cow	Lea	Car	Fac	Air	Mot	Fac	Car
[Serre et al, 2005]	**Model [Serre et al, 2005]**			**97.0**	**99.7**	**98.2**	**96.7**	**98.0**	**95.9**	**95.1**
	Constellation [Weber et al, 2000, Fergus et al, 2003]			84.0	84.8	96.4	94.0	95.0		
	Component-based [Heisele et al, 2002]								90.4	
	Component-based [Leung, 2004]									75.4
[Chikkerur & Wolf, 2006]	**Model [Serre et al, 2005]**	**100.0**	**92.0**	**97.9**		**94.5**		**96.5**		
	Fragments [Epshtein & Ullman, 2005]	98.0	78.7	87.4		66.8		52.6		
	Single template SVM	100.0	77.3	71.6		62.2		65.6		

		MIT-CBCL Street Scene Database						
		Bik	Ped	Car	Bui	Tre	Roa	Sky
[Bileschi & Wolf, 2005]	**Model [Serre et al, 2005]**	**87.8 84.1**	**81.7 88.8**	**89.6 92.9**	**80.3**	**90.8**	**88.9**	**94.7**
	Component-based [Torralba et al, 2004]	68.5	79.8	69.9				
	Part-based [Leibe et al, 2004]	80.9	85.2	85.9				
	Single template SVM	67.8	70.0	85.0				
	Blobworld [Carson et al, 1999]				66.1	85.8	73.1	68.2
	Texton [Renninger & Malik, 2002]				69.7	70.4	58.1	65.1
	Histogram of edges				63.3	63.7	73.3	68.3

Table B.1: Summary of the comparisons performed between the model and other computer vision systems.

Bibliography

H.D.I. Abarbanel, R. Huerta, and M.I. Rabinovich. Dynamical model of long-term synaptic plasticity. *Proc. Nat. Acad. Sci. USA*, 99(15):10132–10137, 2002.

E.H. Adelson and J.R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, 2(2):284–299, 1985.

E.H. Adelson and J.R. Bergen. *Computational models of visual processing*, chapter The plenoptic function and the elements of early vision, pages 3–20. MIT Press, Cambridge, MA, 1991.

T.D. Albright and C.G. Gross. Do inferior temporal cortex neurons encode shape by acting as fourier descriptor filters? In *Proc. of International Conference on Fuzzy Logic and Neural Networks*, pages 375–378, Izuka, Japan, 1990.

B. Amirikian and A.P. Georgopoulos. Modular organization of directionally tuned cells in the motor cortex: is there a short-range order? *Proc. Nat. Acad. Sci. USA*, 100(12474-12479), 2003.

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

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

J.J. Atick. Could information theory provide an ecological theory of sensory processing. *Network: Comput. Neural Syst.*, 3:213–251, 1992.

F. Attneave. Some informational aspects of visual perception. *Psychol. Rev.*, 61:183–193, 1954.

N. Bacon-Mace, M.J. Mace, M. Fabre-Thorpe, and S.J. Thorpe. The time course of visual processing: backward masking and natural scene categorisation. *Vis. Res.*, 45:1459–1469, 2005.

C.I. Baker, M. Behrmann, and C.R. Olson. Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nat. Neurosci.*, 5:1210–1216, 2002.

C. Baleydier and A. Morel. Segregated thalamocortical pathways to inferior parietal and inferotemporal cortex in macaque monkey. *Vis. Neurosci.*, 8:391–405, 1992.

H.B. Barlow. *Sensory Communication*, chapter Possible principles underlying the transformation of sensory messages, pages 217–234. MIT Press, Cambridge, MA, WA Rosenblith edition, 1961.

H.B. Barlow and W.R. Lewick. The mechanism of directionally selective units in rabbit's retina. *J. Neurophys.*, 178(3):477–704, 1965.

G.C. Baylis, E.T. Rolls, and C.M. Leonard. Selectivity between faces in the responses of a population of neurons in the cortex of superior temporal sulcus of the macaque monkey. *Brain Res.*, 342:91–102, 1985.

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

A.C. Berg, T.L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognit.*, 2005.

P. Berkes and L. Wiskott. Slow feature analysis yields a rich repertoire of complex cell properties. *J. Vision*, 5(6):579–602, 2005.

D.N. Bhat and S.K. Nayar. Ordinal measures for image correspondence. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(4):415–423, 1998.

G.Q. Bi and M.M. Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.*, 18: 10464–10472, 1998.

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

I. Biederman. Perceiving real-world scenes. *Science*, 177(43):77–80, 1972.

I. Biederman, J.C. Rabinowitz, A.L. Glass, and E.W. Stacy. On the information extracted from a glance at a scene. *J. Exp. Psych.*, 103(3):597–600, 1974.

S. Bileschi and L. Wolf. Image representations beyond histograms of orientations: Nonlinear gestalt descriptors. 2006. In submission.

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

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

L.J. Borg-Graham and Y. Fregnac. Visual input evokes transient and strong shunting inhibition in visual cortical neurons. *Nature*, 393:369–373, 1998.

A. Borst, V.L. Flanagin, and H. Sompolinsky. Adaptation without parameter change: dynamic gain control in reichardt motion detectors. *Proc. Nat. Acad. Sci. USA*, 102(17): 6172–6176, 2005.

D. Boussaoud, L. G. Ungerleider, and R. Desimone. Pathways for motion analysis: cortical connections of the medialsuperior temporal and fundus of the superior temporal visual areas in the macaque. *J. Comp. Neurol.*, 296(3):462–95, June 1990.

D. Boussaoud, R. Desimone, and L. G. Ungerleider. Visual topography of area TEO in the macaque. *J. Comp. Neurol.*, 306(4):554–75, April 1991.

G.M. Boynton and J. Hegdé. Visual cortex: The continuing puzzle of area V2. *Curr. Biol.*, 14:523–524, 2004.

D.H. Brainard. The psychophysics toolbox. *Spat. Vis.*, 10:433–436, 1997.

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

C. Bruce, R. Desimone, and C. G. Gross. Visual properties of neurons in a polysensory area in superiortemporal sulcus of the macaque. *J. Neurophys.*, 46(2):369–84, August 1981.

R.L. Buckner and M.E. Wheeler. The cognitive neuroscience of remembering. *Nat. Rev. Neurosci.*, 2:624–634, 2001.

E. A. Buffalo, G. Bertini, L.G. Ungerleider, and R. Desimone. Impaired filtering of distracter stimuli by TE neurons following V4 and TEO lesions in macaques. *Cereb. Cortex*, 15:141–151, 2005.

J. Bullier. Integrated model of visual processing. *Brain Res. Rev.*, 36, 2001.

J. Bullier, J.M. Hupe, A. James, and P. Girard. Functional interactions between areas V1 and V2 in the monkey. *J. Phys.*, 90:217–220, 1996.

A. Burkhalter and D. C. Van Essen. Processing of color, form and disparity information in visual areasVP and V2 of ventral extrastriate cortex in the macaque monkey. *J. Neurosci.*, 6(8):2327–51, August 1986.

M. C. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *European Conference on Computer Vision (ECCV)*, pages 628–641, 1998.

C. Cadieu. Modeling shape representation in visual cortex area v4. Master's thesis, MIT, EECS, 2005.

E.M. Callaway. Local circuits in primary visual cortex of the macaque monkey. *Ann. Rev. Neurosci.*, 21:47–74, 1998a.

E.M. Callaway. Visual scenes and cortical neurons: What you see is what you get. *Proc. Nat. Acad. Sci. USA*, 95(7):3344–3345, 1998b.

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

M. Carandini, J.B. Demb, V. Mante, D.J. Tolhurst, Y. Dan, B.A. Olshausen, J.L. Gallant, and N.C. Rust. Do we know what the early visual system does? *J. Neurosci.*, 25(46): 10577–10597, 2005.

C. Carson, M. Thomas, S. Belongie, J. Hellerstein, and J. Malik. Blobworld: A system for region-based image indexing and retrieval. In *Third International Conference on Visual Information Systems*, 1999.

A. Casile and M. Giese. Critical features for the recognition of biological motion. *J. Vision*, 5:348–360, 2005.

S. Celebrini, S. Thorpe, Y. Trotter, and M. Imbert. Dynamics of orientation coding in area V1 of the awake primate. *Vis. Neurosci.*, 10(5):811–25, September 1993.

T. Chi, P. Ru, and S.A. Shamma. Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.*, 118(2):887–906, 2005.

S. Chikkerur and L. Wolf. Empirical comparison between hierarchical fragments based and standard model based object recognition systems. CBCL Paper MMVI-0I, MIT, 2006.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognit.*, 2005.

C. M. Christoudias, B. Georgescu, and P. Meer. Synergism in low level vision. In *Proc. of the Intern. Conf. Comput. Vision*, volume IV, pages 150–155, August 2002.

C. E. Connor, D. C. Preddie, J. L. Gallant, and D. C. VanEssen. Spatial attention effects in macaque area V4. *J. Neurosci.*, 17(9):3201–14, May 1997.

D.M. Coppola, H.R. Purves, A.N. McCoy, and D. Purves. The distribution of oriented contours in the real world. *Proc. Nat. Acad. Sci. USA*, 95(7):4002–4006, 1998.

E. Corthout, B. Uttl, V. Walsh, M. Hallett, and A. Cowey. Timing of activity in early visual cortex as revealed by transcranial magnetic stimulation. *Neuroreport*, 1999.

R. E. Crist, W. Li, and C. D. Gilbert. Learning to see: Experience and attention in primary visual cortex. *Nat. Neurosci.*, 4:519–525, 2001.

G. Csurka, C.R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proc. of the ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.

J.G. Daugman. Two-dimensional spectral analysis of cortical receptive field profile. *Vis. Res.*, 20:847–856, 1980a.

J.G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A*, 2:847–856, 1980b.

P. Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematcial Modeling of Neural Systems*. MIT Press, 2001.

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

G.C. DeAngelis, A. Anzai, I. Ohzawa, and R.D. Freeman. Receptive field structure in the visual cortex: Does selective stimulation induce plasticity? *Proc. Nat. Acad. Sci. USA*, 92: 9682–9686, 1995.

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

A. Delorme and S.J. Thorpe. Face identification using one spike per neuron: resistance to image degradations. *Neural Netw.*, 14:795–803, 2001.

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

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

R. Desimone and C.G. Gross. Visual areas in the temporal lobe of the macaque. *Brain Res.*, 178:363–380, 1979.

R. Desimone, J. Fleming, and C.D. Gross. Prestriate afferents to inferior temporal cortex: an hrp study. *Brain Res.*, 184:41–55, 1980.

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

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

R.L. DeValois, D.G. Albrecht, and L.G. Thorell. Spatial frequency selectivity of cells in macaque visual cortex. *Vis. Res.*, 22:545–559, 1982a.

R.L. DeValois, E.W. Yund, and N. Hepler. The orientation and direction selectivity of cells in macaque visual cortex. *Vis. Res.*, 22:531–544, 1982b.

L. Devroye, G. Laszlo, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer-Verlag, New York, 1996.

E. A. DeYoe and D. C. Van Essen. Concurrent processing streams in monkey visual cortex. *Trends. Neurosci.*, 11(5):219–26, May 1988.

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

R.J. Dolan, G.R. Fink, E. Rolls, M. Booth, A. Holmes, R.S. Frackowiak, and K.J. Friston. How the brain learns to see objects and faces in an impoverished context. *Nature*, 389 (6651):596–599, 1997.

R. J. Douglas and K. A. Martin. A functional microcircuit for cat visual cortex. *J. Physiol. (Lond).*, 440:735–69, 507 1991.

P. E. Downing, Y. Jiang, M. Shuman, and N. Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293:2470–2473, 2001.

M. Egelhaaf and W. Reichardt. Dynamic response properties of movement detectors: Theoretical analysis and electrophysiological investigation in the visual system of the fly. *Biol. Cyb.*, 56:69–87, 1987.

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

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

G.N. Elston. *The primate visual system*, chapter Comparative studies of pyramidal neurons in visual cortex of monkeys, pages 365–385. CRC Press., Boca Raton, FL, 2003.

B. Epshtein and S. Ullman. Feature hierarchies for object classification. In *Proc. of the Intern. Conf. Comput. Vision*, pages 220–227, 2005.

R. Epstein and N. Kanwisher. A cortical representation of the local visual environment. *Nature*, 392:598–601, 1998.

C.A. Erickson and R. Desimone. Responses of macaque perirhinal neurons during and after visual stimulus association learning. *J. Neurosci.*, 19:10404–10416, 1999.

C.A. Erickson, B. Jagadeesh, and R. Desimone. Clustering of perirhinal neurons with similar properties following visual experience in adult monkeys. *Nat. Neurosci.*, 3:1143–1148, 2000.

B.A. Eriksen and C.W. Eriksen. Effects of noise letters upon the identification of a target letter in a non-search task. *Percept. Psychophys.*, 16:143–149, 1974.

K.K Evans and A. Treisman. Perception of objects in natural scenes: Is it really attention free? *J. Exp. Psych.: Hum. Percept. Perf.*, 31(6):1476–1492, 2005.

M. Fabre-Thorpe, G. Richard, and S.J. Thorpe. Rapid categorization of natural images by rhesus monkeys. *Neuroreport*, 9(2):303–308, 1998.

M. Fabre-Thorpe, A. Delorme, C. Marlot, and S.J. Thorpe. A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *J. Comp. Neurol.*, 13(2):171–180, 2003.

F.L. Fahy, I.P. Riches, and M.W. Brown. Neuronal activity related to visual recognition memory: long-term memory and the encoding of recency and familiarity information in the primate anterior and medial inferior temporal and rhinal cortex. *Exp. Brain Res.*, 96: 457–472, 1993.

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

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

D.J. Felleman, A. Burkhalter, and D.C. Van Essen. Cortical connections of areas V3 and VP of macaque monkeyextrastriate visual cortex. *J. Comp. Neurol.*, 379(1):21–47, March 1997.

R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognit.*, volume 2, pages 264–271, 2003.

D. Ferster and K.D. Miller. Neural mechanisms of orientation selectivity in the visual cortex. *Ann. Rev. Neurosci.*, 23:441–471, 2000.

P.C. Fletcher, C.D. Frith, S.C. Baker, T. Shallice, R.S. Frackowiak, and R.J. Dolan. The mind's eye – activation of the precuneus in memory related imagery. *Neuroimage*, 2:196–200, 1995.

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

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

P. Földiák and M.P. Young. *The handbook of brain theory and neural networks*, chapter Sparse coding in the primate cortex, pages 895–898. MIT Press, Cambridge, 1995.

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

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

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

D.J. Freedman, M. Riesenhuber, T. Poggio, and E.K. Miller. Experience dependent sharpening of visual shape selectivity in inferior temporal cortex. *Cereb. Cortex*, 2006. in press.

I. Fujita, K. Tanaka, M. Ito, and K. Cheng. Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, 360:343–346, 1992.

K. Fukushima. Cognitron: A self-organizing multilayered neural network. *Biol. Cyb.*, 20 (3-4):121–136, 1975.

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

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

D. Gabor. Theory of communication. *J. IEE*, 93:429–459, 1946.

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

R. Gattass, A.P. Sousa, M. Mishkin, and L.G. Ungerleider. Cortical projections of area V2 in the macaque. *Cereb. Cortex*, 7:110–129, 1997.

I. Gauthier, M.J. Tarr, A.W. Anderson, P. Skudlarski, and J.C. Gore. Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nat Neurosci*, 2 (6):568–73, 1999.

J. Gautrais and S.J. Thorpe. Rate coding versus temporal order coding: a theoretical approach. *Biosystems*, 1998.

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

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

K.R. Gegenfurtner, D.C. Kiper, and S.B. Fenstemaker. Processing of color, form, and motion in macaque area V2. *Vis. Neurosci.*, 13:161–172, 1996.

A.P. Georgopoulos, J.F. Kalaska, R.Caminiti, and J.T. Massey. On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J. Neurosci.*, 2:1527–1537, 1982.

G. M. Ghose, T. Yang, and J. H. R. Maunsell. Physiological correlates of perceptual learning in monkey V1 and V2. *J. Neurophys.*, 87:1867–1888, 2002.

M. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements and action. *Nat. Rev. Neurosci.*, 4:179–192, 2003.

P. Girard, J.M. Hupe, and J. Bullier. Feedforward and feedback connections between areas V1 and V2 of the monkey have similar rapid conduction velocities. *J. Neurophysiol.*, 85 (3):1328–31, 2001.

P. M. Gochin. Properties of simulated neurons from a model of primate inferior temporal cortex. *Cereb. Cortex*, 5:532–543, 1994.

K. Grauman and T. Darrell. The pyramid match kernel:discriminative classification with sets of image features. In *Proc. of the Intern. Conf. Comput. Vision*, 2005.

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

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

S.. Grossberg. *Neurobiology of attention*, chapter Linking attention to learning, expectation, competition, and consciousness, pages 652–662. Elsevier, San Diego, 2005.

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

R. Guyonneau, H. Kirchner, and S.J. Thorpe. Animals roll around the clock: The rotation invariance of ultra-rapid visual processing. Eur. Conf. on Visual Proc., 2005.

M.E. Hasselmo, E.T. Rolls, and G.C. Baylis. The role of expression and identity in the face selective response of neurons in the temporal visual cortex of the monkey. *Behav. Brain Res.*, 32:203–218, 1989.

M.J. Hawken and A.J. Parker. Spatial properties of neurons in the monkey striate cortex. *Proc. R. Soc. Lond.*, 231:251–288, 1987.

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

D. J. Heeger. Normalization of cell responses in cat striate cortex. *Vis. Res.*, 9(2):181–97, August 1992a.

D. J. Heeger. Half-squaring in responses of cat striate cells. *Vis. Res.*, 9(5):427–43, November 1992b.

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

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

J. Hegdé and D. C. van Essen. Selectivity for complex shapes in primate visual area V2. *J. Neurosci.*, 20:R61:1–6, 2000.

J. Hegdé and D. C. van Essen. Strategies of shape representation in macaque visual area V2. *Vis. Neurosci.*, 20:313–328, 2003.

B. Heisele, P. Ho, and T. Poggio. Face recognition with support vector machines: Global versus component-based approach. In *Proc. of the Intern. Conf. Comput. Vision*, pages 688–694, Vancouver, Canada, 2001a.

B. Heisele, T. Serre, S. Mukherjee, and T. Poggio. Feature reduction and hierarchy of classifiers for fast object detection in video image. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognit.*, volume 2, pages 18–24, Kauai, 2001b.

B. Heisele, T. Serre, M. Pontil, and T. Poggio. Component-based face detection. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognit.*, volume 1, pages 657–662, Hawaii, 2001c.

B. Heisele, T. Serre, M. Pontil, T. Vetter, and T. Poggio. Categorization by learning and combining object parts. In *Advances in Neural Information Processing Systems*, volume 14, 2002.

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

J.A. Hirsch. Synaptic physiology and receptive field structure in the early visual pathway of the cat. *Cereb. Cortex*, 13:63–69, 2003.

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

A.L. Hodgkin and A.F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Neurophys.*, 117:500–544, 1952.

A.D. Holub, M. Welling, and P. Perona. Combining generative models and fisher kernels for object class recognition. In *Proc. of the Intern. Conf. Comput. Vision*, 2005a.

A.D. Holub, M. Welling, and P. Perona. Exploiting unlabelled data for hybrid object classification. In *Neural Information Processing Systems, Workshop in Inter-Class Transfer*, 2005b.

C.L. Martin-Elkins J.A. Horel. Cortical afferents to behaviorally defined regions of the inferior temporal and parahippocampal gyri as demonstrated by WGA-HRP. *J. Comp. Neurol.*, 321(2):177–192, 1992.

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

D.H. Hubel and T.N. Wiesel. Receptive fields of single neurons in the cat's striate visual cortex. *J. Phys.*, 148:574–591, 1959.

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

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

D.H. Hubel and T.N. Wiesel. Functional architecture of macaque monkey. *Proc. R. Soc. Lond. B Biol. Sci.*, 198:1–59, 1977.

C. Hung, G. Kreiman, T. Poggio, and J. DiCarlo. Fast read-out of object identity from macaque inferior temporal cortex. *Science*, 310:863–866, November 2005.

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

M. Ito and H. Komatsu. Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *J. Neurosci.*, 24:3313–3324, 2004.

B. Jagadeesh, L. Chelazzi, M. Mishkin, and R. Desimone. Learning increases stimulus salience in anterior inferior temporal cortex of the macaque. *J. Neurophys.*, 86:290–303, 2001.

R.S. Johansson and I. Birznieks. First spikes in ensembles of human tactile afferents code complex spatial fingertip events. *Nat. Neurosci.*, 7:170–177, 2004.

J.S. Johnson and B. A. Olshausen. Timecourse of neural signatures of object recognition. *J. Vision*, 3:499–512, 2003.

J.P. Jones and L.A. Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophys.*, 58:1233–1258, 1987.

F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proc. of the Intern. Conf. Comput. Vision*, volume 1, pages 604–610, 2005.

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

N. Kanwisher. *The Visual Neurosciences*, chapter The ventral visual object pathway in humans: Evidence from fMRI, pages 1179–1189. MIT Press, 2003.

N. Kanwisher and E. Wojciulik. Visual attention: Insights from brain imaging. *Ann. Rev. Neurosci.*, 1:91–100, 2000.

N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J. Neurosci.*, 17:4302–4311, 1997.

E.K. Karla and A. Treisman. Perception of objects in natural scenes: Is it really attention free? *J. Exp. Psych.: Hum. Percept. Perf.*, 31(6):1476–1492, 2005.

A. Karni and D. Sagi. Where practice makes perfect in texture discrimination: Evidence for primary visual cortex plasticity. *Proc. Natl. Acad. Sci. USA*, 88:4966–4970, 1991.

D. Kersten and A. Yuille. Bayesian models of object perception. *Curr. Op. Neurobiol.*, 13(2): 1–9, 2003.

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

H. Kirchner and S.J. Thorpe. Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vis. Res.*, 2005.

J.J. Knierim and D.C. van Essen. Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *J. Neurophys.*, 67(4):961–p80, 1992.

D.C. Knill and W. Richards. *Perception as Bayesian Inference*. Cambridge: Cambridge University Press, 1996.

U. Knoblich and M. Riesenhuber. Stimulus simplification and object representation: A modeling study. AI Memo 2002-004 / CBCL Memo 215, MIT AI Lab and CBCL, Cambridge, MA, 2002.

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

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

K.P. Körding, C. Kayser, W. Einhuser, and P. König. How are complex cell properties adapted to the statistics of natural stimuli? *J. Neurophys.*, 91(1):206–212, 2004.

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

G. Kovács, R. Vogels, and G.A. Orban. Cortical correlate of pattern backward masking. *Proc. Nat. Acad. Sci. USA*, 92:5587–5591, 1995.

V.A.F. Lamme and P.R. Roelfsema. The disctinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosci.*, 23:571–579, 2000.

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

S. Lazebnik, C. Schmid, and J. Ponce. A maximum entropy framework for part-based texture and object recognition. In *Proc. of the Intern. Conf. Comput. Vision*, 2005.

Y. LeCun. Learning processes in an asymmetric threshold network. In *Disordered systems and biological organization*, pages 233–240, Les Houches, France, 1986.

Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Comp.*, 1(4):541–551, 1989.

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

Y. LeCun, F. J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognit.* IEEE Press, 2004.

Y. LeCun, U. Muller, J. Ben, E., Cosatto, and B. Flepp. Off-road obstacle avoidance through end-to-end learning. In *Advances in Neural Information Processing Systems*. MIT Press, 2005.

B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *SLCP '04 Workshop on Statistical Learning in Computer Vision*, 2004.

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

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

T.K. Leung, M.C. Burl, and P. Perona. Finding faces in cluttered scenes using random labeled graph matching. In *Proc. of the Intern. Conf. Comput. Vision*, pages 637–644, Cambridge, MA, 1995.

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

L. Li, E.K. Miller, and R. Desimone. The representation of stimulus familiarity in anterior inferiortemporal cortex. *J. Neurophysiol.*, 69(6):1918–29, June 1993.

V. Litvak, H. Sompolinsky, I. Segev, and M. Abeles. On the transmission of rate code in long feedforward networks with excitatory-inhibitory balance. *J. Neurosci.*, 23(7):3006–3015, 2003.

J. Liu, A. Harris, and N. Kanwisher. Stages of processing in face perception: An MEG study. *Nat. Neurosci.*, 5(9):910–916, 2002.

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

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

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

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

D.G. Lowe. Towards a computational model for object recognition in IT cortex. In *Proc. of Biologically Motivated Computer Vision*, pages 20–31, Seoul, Korea, 2000.

D.G. Lowe. Distinctive image features from scale-invariant keypoints. *Intern. J. Comput. Vision*, 2004.

D.G. Lowe. Object recognition from local scale-invariant features. In *Proc. of the Intern. Conf. Comput. Vision*, pages 1150–1157, 1999.

J.S. Lund, A. Angelucci, and P.C. Bressloff. Anatomical substrates for functional columns in macaque monkey primary visual cortex. *Cereb. Cortex*, 12:15–24, 2003.

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

L.E. Mahon and R.L. DeValois. Cartesian and non-cartesian responses in LGN, V1, and V2 cells. *Vis. Neurosci.*, 18:973–981, 2001.

P. Mamassian, M.S. Landy, and L.T. Maloney. *Probabilistic Models of the Brain: Perception and Neural Function*, chapter Bayesian modelling of visual perception, pages 13–36. MIT Press, Cambridge, MA, 2002.

S. Marcelja. Mathematical description of the responses of simple cortical cells. *J. Opt. Soc. Am. A*, 70, 1980.

H. Markram. The blue brain project. *Nat. Rev. Neurosci.*, 7:153–160, 2006.

H. Markram, J. Lübke, M. Frotscher, and B. Sakmann. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275:213–215, 1997.

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

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

D. McLaughlin, R. Shapley, M. Shelley, and D. J. Wielaard. A neuronal network model of macaque primary visual cortex (V1): Orientation selectivity and dynamics in the input layer 4ca. *Proc. Nat. Acad. Sci. USA*, 97(14):8087–8092, 2000.

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

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

W. H. Merigan, T. A. Nealey, and J. H. Maunsell. Visual effects of lesions of cortical area V2 in macaques. *J. Neurosci.*, 13(7):3180–91, July 1993.

M.M. Merzenich and J.F. Brugge. Representation of the cochlear partition of the superior temporal plane of the macaque monkey. *Brain Res.*, 50:275–296, 1973.

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

E.K. Miller, L. Li, and R. Desimone. A neural mechanism for working and recognition memory in inferior temporal cortex. *Science*, 254(5036):1377–9, November 1991.

E.K. Miller, L. Li, and R. Desimone. Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *J. Neurosci.*, 13(4):1460–78, April 1993.

K.D. Miller. Understanding layer 4 of the cortical circuit: A model based on cat V1. *Cereb. Cortex*, 2003.

M. Mishkin, L.G. Ungerleider, and K.A. Macko. Object vision and spatial vision: Two cortical pathways. *Trends Neurosci.*, 1983.

Y. Miyashita. Inferior temporal cortex: Where visual perception meets memory. *Ann. Rev. Neurosci.*, 16:245–263, 1993.

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

Y. Miyashita and T. Hayashi. Neural representation of visual objects: Encoding and top-down activation. *Curr. Op. Neurobiol.*, 10:187–194, 2000.

A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, volume 23, pages 349–361, 2001.

J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate-cortex. *Science*, 229(4715):782–4, August 1985.

Y. Mouchetant-Rostaing, M.H. Giard, C. Delpuech, J.F. Echallier, and J. Pernier. Early signs of visual categorization for biological and non-biological stimuli in humans. *Neuroreport*, 11(11):2521–2525, 2000.

V.B. Mountcastle. Modality and topographic properties of single neurons of cat's somatic sensory cortex. *J. Neurophys.*, 20:408–434, 1957.

V.B. Mountcastle. The columnar organization of the neocortex. *Brain*, 120(Part 4):701–22, 1997.

D. Mumford. On the computational architecture of the neocortex – II: The role of cortico-cortical loops. *Biol. Cyb.*, 66:241–251, 1992.

J. Mutch and D. Lowe. Multiclass object recognition using sparse, localized hmax features. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognit.*, 2006. To appear.

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

J.R. Newton and M. Sur. *Plasticity and signal representation in the auditory system*, chapter Rewiring cortex functional visual plasticity in the auditory cortex during development. Springer, 2004.

L.G. Nowak and J. Bullier. *Extrastriate visual cortex in primates*, volume 12, chapter The timing of information transfer in the visual system, pages 205–241. New York: Plenum Press, 1997.

K. O'Craven, P. Downing, and N. Kanwisher. fMRI evidence for objects as the units of attentional selection. *Nature*, 401:584–587, 1999.

A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Prog. in Brain Res.*, In press.

A. Oliva, A. Torralba, M.S. Castelhano, and J.M. Henderson. Top down control of visual attention in object detection. In *Proc. IEEE Proceedings of the International Conference on Image Processing*, volume 1, pages 253–256, 2003.

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

B.A. Olshausen, C.H. Anderson, and D.C. Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.*, 13(11):4700–19, November 1993.

B.A. Olshausen, C.H. Anderson, and D.C. Van Essen. A multiscale dynamic routing circuit for forming size- and position-invariant object representations. *J. Comput. Neurosci.*, 2(1): 45–62, March 1995.

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

H. op de Beeck, H. Wagemans, and R. Vogels. The effect of category learning on the representation of shape: Dimensions can be biased but not differentiated. *J. Exp. Psychol. Gen.*, 132:491–511, 2003.

M.W. Oram and D.I. Perrett. Time course of neural responses discriminating different views of the face and head. *J. Neurophys.*, 68:70–84, 1992.

M.W. Oram and D.I. Perrett. Modeling visual recognition from neurobiological constraints. *Neural Netw.*, 7(6-7):945–972, 1994.

M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognit.*, pages 193–199, San Juan, 1997.

E. Osuna. *Support Vector Machines: Training and Applications*. PhD thesis, MIT, Department of Electrical Engineering and Computer Science, Cambridge, MA, 1998.

E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognit.*, pages 130–136, 1997.

D.B. Parker. *Neural Networks for Computing*, chapter A comparison of algorithms for neuron-like cells, pages 327–332. American Institute of Physics, New York, 1986.

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

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

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

A. Pasupathy and E.K. Miller. Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*, 24:873–876, 2005.

D.G. Pelli. The video toolbox software for visual psychophysics: Transforming numbers into movies. *Spat. Vis.*, 1997.

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

D.I. Perrett, E.T. Rolls, and W. Caan. Visual neurones responsive to faces in the monkey temporal cortex. *Exp. Brain Res.*, 47(3):329–42, 1982.

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

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

D.I. Perrett, A.J. Mistlin, and A.J. Chitty. Visual neurones responsive to faces. *Trends in Neuroscience*, 10:358–364, 1987.

D.I. Perrett, M.W. Oram, M.H. Harries, R. Bevan, J.K. Hietanen, P.J. Benson, and S. Thomas. Viewer-centred and object-centred coding of heads in the macaque temporal cortex. *Exp. Brain Res.*, 86:159–173, 1991.

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

D.I. Perrett, M.W. Oram, and E. Wachsmuth. Evidence accumulation in cell populations responsive to faces: An account of generalisation of recognition without mental transformations. *Cognition*, 67:111–145, 1998.

T. Poggio. A theory of how the brain might work. In *Proc. of Cold Spring Harbor Symposia on Quantitative Biology*, volume 4, pages 899–910, Cold Spring Harbor, NY, 1990. Cold Spring Harbor Laboratory Press.

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

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

T. Poggio and F. Girosi. Networks for approximation and learning. *Proc. IEEE*, 78(9), 1990.

T. Poggio and A. Hurlbert. *Large Scale Neuronal Theories of the Brain*, chapter Observations on Cortical Mechanisms for Object Recognition and Learning, pages 153–182. MIT press, Cambridge, MA, 1994.

T. Poggio and S. Smale. The mathematics of learning: Dealing with data. *Notices of the american Mathematical Society (AMS)*, 50(5), 2003.

E.O. Postma, H.J. van den Herik, and P.T.W. Hudson. Scan: A scalable neural model of covert attention. *Nat. Neurosci.*, 10(6):993–1015, 1997.

M.C. Potter. Short-term conceptual memory for pictures. *J. Exp. Psych.: Hum. Learn. Mem.*, 2:509–522, 1976.

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

M.C. Potter, A. Staub, J. Rado, and D.H. O'Connor. Recognition memory for briefly presented pictures: The time course of rapid forgetting. *J. Exp. Psych.: Hum. Percept. Perf.*, 28(5):1163–1175, 2002.

G. Rainer and E.K. Miller. Effects of visual experience on the representation of objects in the prefrontal cortex. *Neuron*, 27:8–10, 2000.

G. Rainer, H. Lee, and N.K. Logothetis. The effect of learning on the function of monkey extrastriate visual cortex. *PLoS Biology*, pages 275–284, 2004.

R. Raizada and S. Grossberg. Context-sensitive bindings by the laminar circuits of V1 and V2: A unified model of perceptual grouping, attention, and orientation contrast. *Vis. Cognition*, 8:431–466, 2001.

R.N. Rao, B.A. Olshausen, and M.S. Lewicki. *Probabilistic Models of the Brain: Perception and Neural Function*. MIT Press, Cambridge, MA, 2002.

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

W. Reichardt. *Sensory Communication*, pages 303–317. MIT Press, Cambridge, MA, 1961.

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

L.W. Renninger and J. Malik. When is scene identification just texture recognition? *Vis. Res.*, 44:2301–2311, 2004.

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

W. Richards, J. Feldman, and A. Jepson. From features to perceptual categories. In *Proc. of British Machine Vision Conf.*, pages 99–108, 1992.

M. Riesenhuber and P. Dayan. Neural models for part-whole hierarchies. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, pages 17–23, Cambridge, MA, 1997. MIT Press.

M. Riesenhuber and T. Poggio. Neural mechanisms of object recognition. *Curr. Op. Neurobiol.*, 12:162–168, 2002.

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

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, 2:1019–1025, 1999a.

M. Riesenhuber and T. Poggio. Are cortical models really bound by the “Binding Problem”? *Neuron*, 24:87–93, 1999b.

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

D.L. Ringach, M.J. Hawken, and R. Shapley. Dynamics of orientation tuning in macaque primary visual cortex. *Nature*, 387(6630):281–4, May 1997.

K.S. Rockland. Visual cortical organization at the single axon level: A beginning. *Neurosci. Res.*, 42(3):155–66, March 2002.

K.S. Rockland and A. Virga. Organization of individual cortical axons projecting from area V1 (area 17) to V2 (area 18) in the macaque monkey. *Vis. Neurosci*, 4(1):11–28, January 1990.

K.S. Rockland, K.S. Saleem, and K. Tanaka. Divergent feedback connections from areas V4 and TEO in the macaque. *Vis. Neurosci.*, 11(3):579–600, 1994.

E. T. Rolls. Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. *Hum. Neurobiol.*, 3(4):209–22, 1984.

E.T. Rolls. *The Visual Neurosciences*, chapter Invariant object and face recognition, pages 1165–1178. MIT Press, Cambridge, MA, 2004.

E.T. Rolls. The orbitofrontal cortex and reward. *Cereb. Cortex*, 10:284–294, 2000.

E.T. Rolls. Learning mechanisms in the temporal lobe visual cortex. *Behav. Brain Res.*, 66 (1-2):177–185, 1995.

E.T. Rolls and T. Milward. A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Comp.*, 12(11):2547–2572, 2000.

E.T. Rolls and M.J. Tovee. Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proc. R. Soc. Lond. B Biol. Sci.*, 1994.

E.T. Rolls, A. Coway, and V. Bruce. Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philos. Trans. Roy. Soc*, 335:11–21, 1992.

E.T. Rolls, M.J. Tovee, and S. Panzeri. The neurophysiology of backward visual masking: Information analysis. *J. Comp. Neurol.*, 11:300–311, 1999.

F. Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan books, Washington D.C., 1962.

N.P. Rougier and R.C. O' Reilly. A gated prefrontal cortex model of dynamic task switching. *Cognitive Science*, 26:503–520, 2002.

N.P. Rougier, D.C. Noelle, T.S. Braver, J.D. Cohen, and R.C. O' Reilly. Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proc. Nat. Acad. Sci. USA*, 102(20): 7338–7343, 2005.

G.A. Rousselet, M. Fabre-Thorpe, and S.J. Thorpe. Parallel processing in high level categorisation of natural images. *Nat. Neurosci.*, 5:629–630, 2002.

G.A. Rousselet, M.J. Mace, and M. Fabre-Thorpe. Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *J. Vision*, 3:440–455, 2003.

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

G.A. Rousselet, S.J. Thorpe, and M. Fabre-Thorpe. Processing of one, two or four natural scenes in humans: the limits of parallelism. *Vis. Res.*, 44:877–894, 2004b.

D. Ruderman. The statistics of natural images. *Network : Computation in Neural Systems*, 5: 598–605, 1994.

D.E. Rumelhart, G.E. Hinton, and R.J. Williams. *Parallel distributed processing: explorations in the microstructure of cognition*, volume 1: foundations, chapter Learning internal representations by error propagation, pages 318–362. MIT Press, Cambridge, 1986.

J. Sadr, S. Mukherjee, K. Thoresz, and P. Sinha. The fidelity of local ordinal encoding. In S. Becker T. Dietterich and Z. Ghahramani, editors, *Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

K. Sakai and Y. Miyashita. Neural organization for the long-term memory of paired associates. *Nature*, 354:152–155, 1991.

K. Sakai and Y. Miyashita. Neuronal tuning to learned complex forms in vision. *Neuroreport*, 5:829–832, 1994.

K.S. Saleem, K. Tanaka, and K.S. Rockland. Pha-l study of connections from TEO and V4 to TE in the monkey visual cortex. *Society for Neuroscience Abstracts*, 18(294), 1992.

K.S Saleem, K. Tanaka, and K.S. Rockland. Specific and columnar projection from area TEO to TE in the macaque inferotemporal cortex. *Cereb. Cortex*, 3:54–64, 1993.

K.S. Saleem, W. Suzuki, K. Tanaka, and T. Hashikawa. Connections between anterior inferotemporal cortex and superior temporal sulcus regions in the macaque monkey. *J. Neurosci.*, 20:5083–5101, 1996.

E. Sali and S. Ullman. Combining class-specific fragments for object classification. In *Proc. of British Machine Vision Conf.*, 1999.

P.A. Salin and J. Bullier. Corticocortical connections in the visual system: structure and function. *Physiol. Rev.*, 75:107–154, 1995.

E. Salinas and L. F. Abbott. Invariant visual responses from attentional gain fields. *J. Neurophys.*, 77:3267–3272, 1997.

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

S.P.O. Scalaidhe, F.A.W. Wilson, and P.S. Goldman-Rakic. Areal segregation of face-processing neurons in prefrontal cortex. *Science*, 278(5340):1135–1138, 1999.

P. H. Schiller, B. L. Finlay, and S. F. Volman. Quantitative studies of single-cell properties in monkey striate cortex V. Multivariate statistical analyses and models. *J. Neurophysiol.*, 39(6):1362–1374, 1976a.

P. H. Schiller, B. L. Finlay, and S. F. Volman. Quantitative studies of single-cell properties in monkey striate cortex IV. Corticotectal cells. *J. Neurophysiol.*, 39(6):1352–1361, 1976b.

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

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

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

P.H. Schiller. Effect of lesions in visual cortical area V4 on the recognition of transformed objects. *Nature*, pages 342–344, 1995.

P.H. Schiller. The effects of V4 and middle temporal (MT) area lesions on visual performance in the rhesus monkey. *Vis. Neurosci.*, pages 717–746, 1993.

P.H. Schiller and K. Lee. The role of the primate extrastriate area V4 in vision. *Science*, pages 1251–1253, 1991.

G. Schneider, H. Wersing, B. Sendhoff, and E. Körner. Evolutionary optimization of a hierarchical object recognition model. *IEEE Trans. Systems, Man, Cybernetics, Part B: Cybernetics*, 35:426–437, 2005.

H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognit.*, pages 746–751, 2000.

A. Schoups, R. Vogels, N. Qian, and G. Orban. Practising orientation identification improves orientation coding in V1 neurons. *Nature*, 412:549–553, 2001.

S. Schuett, T. Bonhoeffer, and M. Hubener. Pairing-induced changes of orientation maps in cat visual cortex. *Neuron*, 32:325–337, 2001.

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

P. Schyns, R.L. Goldstone, and J.P. Thilbaut. The development of features in object concepts. *Behav. Brain Sci.*, 21:1–54, 1998.

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

T. Serre and T. Poggio. Standard model v2.0: How visual cortex might learn a dictionary of shape-components. J. Vision, 2004.

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

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

T. Serre, T. Poggio, and P. Sihna. Face detection by humans and machines. J. Vision, 2004a.

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

T. Serre, M. Kouh., C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. AI Memo 2005-036 / CBCL Memo 259, MIT, Cambridge, MA, 2005a.

T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In IEEE Computer Society Press, editor, *Proc. IEEE Conf. on Comput. Vision and Pattern Recognit.*, San Diego, 2005b.

T. Serre, A. Oliva, and T. Poggio. A feedforward theory of visual cortex account for human performance in rapid categorization. *in prep.*, 2006a.

T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2006b. Accepted for publication.

H.S. Seung. Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron*, 40:1063–1073, 2003.

M.N. Shadlen and W.T. Newsome. The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. *J. Neurosci.*, 18(10):3870–3896, 1998.

D.L. Sheinberg and N.K. Logothetis. Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. *J. Neurosci.*, 21(4):1340–1350, 2001.

S. Shipp and S. Zeki. The organization of connections between areas v5 and V1 in macaque monkey visual cortex. *Eur. J. Neurosci.*, 1:332–354, 1989a.

S. Shipp and S. Zeki. The organization of connections between areas v5 and V2 in macaque monkey visual cortex. *Eur. J. Neurosci.*, 1:332–354, 1989b.

N. Sigala and N. Logothetis. Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, 415:318–320, 2002.

R. Sigala, T. Serre, T. Poggio, and M. Giese. Learning features of intermediate complexity for the recognition of biological motion. In *Proc. of the Intern. Conf. Artif. Neural Ntw.*, 2005.

A.M. Sillito. *Functional properties of cortical cells*, volume 2, chapter Functional considerations of the operation of GABAergic inhibitory processes in the visual cortex, pages 91–117. New York: Plenum Press, 1984.

A.M. Sillito, H.E. Jones, G.L. Gerstein, and D.C. West. Feature-linked synchronization of thalamic relay cell firing induced by feedback from the visual cortex. *Nature*, 369:479–482, 1994.

T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) database of human faces. Computer Science Technical Report 01-02, CMU, 2001.

E.P. Simoncelli and D.J. Heeger. A model of neural responses in visual area MT. *Vis. Res.*, 38:743–761, 1998.

E.P. Simoncelli and B.A. Olshausen. Natural image statistics and neural representation. *Ann. Rev. Neurosci.*, 24:1193–1216, 2001.

W. Singer, F. Tretter, and U. Yinon. Evidence for long-term functional plasticity in the visual cortex of adult cats. *J. Neurophys.*, 324:239–248, 1982.

P. Sinha. Qualitative representations for recognition. In *Biologically Motivated Computer Vision*, pages 249–262, 2002.

E.C. Smith and M.S. Lewicki. Efficient auditory coding. *Nature*, 2006.

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

H. Sompolinsky and R. Shapley. New perspectives on mechanisms for orientation selectivity. *Curr. Op. Neurobiol.*, 7:514–522, 1997.

W.C. De Souza, S. Eifuku, R. Tamura, H. Nishijo, and T. Ono. Differential characteristics of face neuron responses within the anterior superior temporal sulcus of macaques. *J. Neurophys.*, 94:1252–1266, 2005.

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

R.E. Steele and G.E. Weller. Qualitative and quantitative features of axons projecting from caudal to rostral inferior temporal cortex of squirrel monkeys. *Vis. Neurosci.*, 12(4):701–22, 1995.

C.F. Stevens. Models are common; good theories are scarce. *Nat. Neurosci.*, 3:1177, 2000.

S.M. Stringer and E.T. Rolls. Position invariant recognition in the visual system with cluttered environments. *Neural Netw.*, 13:305–315, 2000.

S.M. Stringer and E.T. Rolls. Invariant object recognition in the visual system with novel views of 3d objects. *Neural Comp.*, 14:2585–2596, 2002.

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

Z. Sun, T. Tan, and Y. Wang. Robust encoding of local ordinal measures: A general framework of iris recognition. In *ECCV workshop on Biometric Authentication*, pages 270 – 282, 2004.

K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.

K.-K. Sung. *Learning and Example Selection for Object and Pattern Recognition*. PhD thesis, MIT, Artificial Intelligence Laboratory and Center for Biological and Computational Learning, Cambridge, MA, 1996.

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

K. Tanaka. Neuronal mechanisms of object recognition. *Science*, 262(5134):685–8., 1993.

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

K. Tanaka. Columns for complex visual object features in the inferotemporal cortex: Clustering of cells with similar but slightly different stimulus selectivities. *Cereb. Cortex*, 13: 90–99, 2003.

K. Tanaka. Mechanisms of visual object recognition: monkey and human studies. *Curr. Op. Neurobiol.*, 7:523–529, 1997.

S.J. Thorpe. Ultra-rapid scene categorisation with a wave of spikes. In *BMCV*, 2002.

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

S.J. Thorpe and J. Gautrais. Rapid visual processing using spike asynchrony. In *Neural Information Processing Systems*, pages 901–907, 1997.

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

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

S.J. Thorpe, A. Delorme, and R. VanRullen. Spike-based strategies for rapid processing. *Neural Netw.*, 14:715–725, 2001a.

S.J. Thorpe, K.R. Gegenfurtner, M. Fabre-Thorpe, and H.H. Bulthoff. Detection of animals in natural images using far peripheral vision. *Eur. J. Neurosci.*, 14:869–876, 2001b.

D. J. Tolhurst and D. J. Heeger. Contrast normalization and a linear model for the directional selectivity of simple cells in cat striate cortex. *Vis. Res.*, 14(1):19–25, January 1997.

A. Torralba and A. Oliva. Depth estimation from image structure. *IEEE Pattern Analysis and Machine Intelligence*, 24, 2002.

A. Torralba and A. Oliva. Statistics of natural image categories. *Network: Computat. Neural Syst.*, 14:391–412, 2003.

A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features: Efficient boosting procedures for multiclass object detection. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognit.*, 2004.

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

M.J. Tovee. Neuronal processing. how fast is the speed of thought? *Curr. Biol.*, 1994.

M.J. Tovee, E.T. Rolls, A. Treves, and R.P. Bellis. Information encoding and the response of single neurons in the primate temporal visual cortex. *J. Neurophys.*, 1993.

A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cog. Psych.*, 12: 97–136, 1980.

L. Tremblay and W. Wolfram. Modifications of reward expectation-related neuronal activity during learning in primate orbitofrontal cortex. *J. Neurophysiol.*, 83, 2000.

S. Ullman and S. Soloviev. Computation of pattern invariance in brain-like structures. *Neural Netw.*, 12:1021–36, 1999.

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

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

M.C.W. van Rossum, G.Q. Bi, and G.G. Turrigiano. Stable hebbian learning from spike timing-dependent plasticity. *J. Neurosci.*, 20(23):8812–8821, 2000.

M.C.W. van Rossum, G.G. Turrigiano, and S.B. Nelson. Fast propagation of firing rates through layered networks of noisy neurons. *J. Neurosci.*, 22:1956–1966, 2002.

R. VanRullen and C. Koch. Visual selective behavior can be triggered by a feed-forward process. *J. Comp. Neurol.*, 15:209–217, 2003.

R. VanRullen and S.J. Thorpe. The time course of visual processing: From early perception to decision-making. *J. Comp. Neurol.*, 13:454–461, 2001a.

R. VanRullen and S.J. Thorpe. Is it a bird? Is it a plane? Ultra-rapid visual characterisation of natural and artifactual objects. *Perception*, 30:655–668, 2001b.

R. VanRullen, J. Gautrais, A. Delorme, and S.J. Thorpe. Face processing using one spike per neurone. *Biosystems*, 48:229–39, 1998.

R. VanRullen, R. Guyonneau, and S.J. Thorpe. Spike times make sense. *Trends in Neurosci.*, 28(1), 2005.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, Berlin, 1995.

T. Vetter, A. Hurlbert, and T. Poggio. View-based models of 3D object recognition: invariance to imaging transformations. *Cereb. Cortex*, 3:261–269, 1995.

J.D. Victor, F. Mechler, M.A. Repucci, K.P. Purpura, and T. Sharpee. Responses of V1 neurons to two-dimensional hermite functions. *J. Neurophys.*, 95:379–400, 2006.

P. Viola and M. Jones. Robust real-time face detection. In *Proc. of the Intern. Conf. Comput. Vision*, volume 20(11), pages 1254–1259, 2001.

R. Vogels. Categorization of complex visual images by rhesus monkeys. Part 2: Single-cell study. *Eur. J. Neurosci.*, 11:1239–1255, 1999.

C. von der Malsburg. Binding in models of perception and brain function. *Curr. Op. Neurobiol.*, 5:520–526, 1995.

C. von der Malsburg. The what and why of binding: The modeler's perspective. *Neuron*, 24:95–125, 1999.

C. von der Malsburg. The correlation theory of brain function. Technical Report Technical Report 81-2, Dept. of Neurobiology, Max-Planck Institute for Biophysical Chemistry, Göttingen, Germany, 1981.

L. von Melchner, S.L. Pallas, and M. Sur. Visual behaviour mediated by retinal projections directed to the auditory pathway. *Nature*, 2000.

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

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

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

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

D. Walther, T. Serre, T. Poggio, and C. Koch. Modeling feature sharing between object detection and top-down attention. *J. Vision*, 5(8):1041–1041, 9 2005. ISSN 1534-7362. URL `http://journalofvision.org/5/8/1041/`.

G. Wang, K. Tanaka, and M. Tanifuji. Optical imaging of functional organization in the monkey inferotemporal cortex. *Science*, 272:1665–1668, 1996.

G. Wang, M. Tanifuji, and K. Tanaka. Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. *Neurosci. Res.*, 32:33–46, 1998.

A.B. Watson and J.A. Solomon. Model of visual contrast gain control and pattern masking. *J. Opt. Soc. Am. A*, 1997.

M. Weber, W. Welling, and P. Perona. Towards automatic discovery of object categories. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognit.*, 2000a.

M. Weber, W. Welling, and P. Perona. Unsupervised learning of models of recognition. In *Proc. of the European Conference on Computer Vision*, volume 2, pages 1001–108, 2000b.

M.J. Webster, L.G. Ungerleider, and J. Bachevalier. Connections of inferior temporal areas TE and TEO with medial temporal-lobe structures in infant and adult monkeys. *J. Neurosci.*, 11:1095–1116, 1991.

M.J. Webster, J. Bachevalier, and L.G. Ungerleider. Connections of inferior temporal areas TEO and TE with parietal and frontal cortex in macaque monkeys. *Cereb. Cortex*, 4:470–483, 1994a.

M.J. Webster, J. Bachevalier, and L.G. Ungerleider. Subcortical connections of inferior temporal areas TE and TEO in macaque monkey. *J. Comp. Neurol.*, 335:73–91, 1994b.

Y. Weiss, E.P. Simoncelli, and E.H. Adelson. Motion illusions as optimal percepts. *Nat. Neurosci.*, 5(6):598–604, 2002.

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

P. Williams, I. Gauthier, and M.J. Tarr. Feature learning during the acquisition of perceptual expertise. *Behavioral and Brain Sciences*, 21(1):40–41, 1998.

F.A.W. Wilson, S.P.O. Scalaidhe, and P.S. Goldman-Rakic. Functional synergism between putative y-aminobutyrate-containing neurons and pyramidal neurons in prefrontal cortex. *Proc. Nat. Acad. Sci. USA*, 91:4009–4013, 94.

L. Wiskott. *Problems in Systems Neuroscience*, chapter How does our visual system achieve shift and size invariance? J.L. van Hemmen and T.J. Sejnowski. Oxford University Press, 2006.

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

L. Wolf, S. Bileschi, and E. Meyers. Perception strategies in hierarchical vision systems. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognit.*, 2006. To appear.

J.M. Wolfe and S.C. Bennett. Preattentive object files: Shapeless bundles of basic features. *Vis. Res.*, 37:25–44, 1997.

J.M. Wolfe, T.S. Horowitz, N. Kenner, M. Hyle, and N. Vasan. How fast can you change your mind? The speed of top-down guidance in visual search. *Vis. Res.*, 44:1411–1426, 2004.

S. Yamane, S. Kaji, and K. Kawano. What facial features activate face neurons in the inferior temporal cortex of the monkey? *Exp. Brain Res.*, 73:209–214, 1988.

T. Yang and J. H. R. Maunsell. The effect of perceptual learning on neuronal responses in monkey visual area V4. *J. Neurosci.*, 24:1617–1626, 2004.

H. Yao and Y. Dan. Stimulus timing-dependent plasticity in cortical processing of orientation. *Neuron*, 32:315–323, 2001.

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

S. Zeki and S. Shipp. Modular connections between areas V2 and V4 of macaque monkey visual cortex. *Eur. J. Neurosci.*, 1:494–506, 1985.

S. Zeki and S. Shipp. The functional logic of cortical connections. *Nature*, 335:311–317, 1988.

H. Zhou, H. S. Friedman, and R. von der Heydt. Coding of border ownership in monkey visual cortex. *J. Neurosci.*, 20:6594–6611, 2000.