Visual attention and object categorization: from psychophysics to computational models

Robert J. Peters

In partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in
Computation and Neural Systems

California Institute of Technology
Pasadena, California
2004
(Defended 4 June 2004)

Abstract

This thesis is arranged in two main parts. Each part relies an approach using the methods of psychophysics and computational modeling to bring abstract or high-level theories of vision closer to a concrete neurobiological foundation.

The first part addresses the topic of visual object categorization. Previous studies using high-level models categorization have left unresolved issues of neurobiological relevance, including how features are extracted from the image and the role played by memory capacity in categorization performance. We compared the ability of a comprehensive set of models to match the categorization performance of human observers while explicitly accounting for the models' numbers of free parameters. The most successful models did not require a large memory capacity, suggesting that a sparse, abstracted representation of category properties may underlie categorization performance. This type of representation—different from classical prototype abstraction—could also be extracted directly from two-dimensional images via a biologically plausible early vision model, rather than relying on experimenter-imposed features.

The second part addresses visual attention in its bottom-up, stimulus-driven form. Previous research [Parkhurst et al., 2002] showed that a model of bottom-up visual attention can account in part for the spatial positions of locations fixated by humans while free-viewing complex natural and artificial scenes. We used a similar framework to quantify how the predictive ability of such a model may be enhanced by new model components based on several specific mechanisms within the functional architecture of the visual system. These components included richer interactions among orientation-tuned units, both at short-range (for clutter reduction) and at long-range (for contour facilitation). Subjects free-viewed naturalistic

and artificial images while their eye movements were recorded. The resulting fixation locations were compared with the models' predicted salience maps. We found that each new model component was important in attaining a strong quantitative correspondence between model and behavior. Finally, we compared the model predictions with the spatial locations obtained from a task that relied on mouse clicking rather than eye tracking. As these models become more accurate in predicting behaviorally-relevant salient locations, they become useful to a range of applications in computer vision and human-machine interface design.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Preface

The passing of a major milestone in life[1] is always an occasion for reflection, especially when, as one colleague put it, "you have had about *this much* [holds hand about an inch in front of eyes] perspective on your research for the past six years." Indeed the view is a bit different when one takes a step back: here is what I wrote (in part) as I applied for admission to the Computation and Neural Systems program at Caltech in fall of 1997:

> I have always been fascinated by the great mysteries of nature, yet at times I have felt reluctant to single one out for further study, at the expense of being forced to neglect the others. Having realized, however, that one cannot specialize in all the mysteries (except perhaps in kindergarten), and uneasy about the prospect of specializing in nothing at all, I decided upon a compromise. I embarked on a search for my center of intellectual gravity—that specialty that would be reinforced rather than undermined by my interests that formally lie outside the specialty.
>
> I have performed this search by the admittedly haphazard procedure of reading one book, mulling it over, then moving on to what I fell impelled to read next. In the past year, this technique has taken me from (roughly in order) *Ishmael* (Daniel Quinn, 1992) to *Dreams of a Final Theory* (Steven Weinberg, 1992), *Shadows of the Mind* (Roger Penrose, 1992), *A Brief History of Time* (Stephen Hawking, 1988), Conscious-

[1] If completing my Ph.D. doesn't qualify as a major milestone, I have the impending rollover of a significant digit (who shall remain nameless) in my age to use as a backup milestone.

ness Explained (Daniel Dennett, 1991), *The Language Instinct* (Stephen Pinker, 1994), *Darwin's Dangerous Idea* (Daniel Dennett, 1995), *At Home in the Universe* (Stuart Kauffman, 1995), *The Extended Phenotype* (Richard Dawkins, 1982), *The Astonishing Hypothesis* (Francis Crick, 1994), *Elbow Room: The Varieties of Free Will Worth Wanting* (Daniel Dennett, 1984), *The Intentional Stance* (Daniel Dennett, 1987), *Higher Superstitions* (Paul Gross and Norman Levitt, 1994), *The Tao of Physics* (Fritjof Capra, 1975), *The Third Culture* (John Brockman, 1995), *The Embodied Mind* (Francisco Varela, Evan Thompson, and Eleanor Rosch, 1991), *The Quark and the Jaguar* (Murray Gell-Mann, 1994), *The Story of B* (Daniel Quinn, 1996), *The Bell Curve* (Richard Herrnstein and Charles Murray, 1994), *The Mismeasure of Man* (Stephen Jay Gould, 1996), and *Neurophilosophy* (Patricia Churchland, 1986). These sources, along with journal articles, some textbook reading, participation in online discussion groups, and conversations with the world-class professors and peers that surround me at the University of Wisconsin, have led me to that center of intellectual gravity that has always been within me but only now apparent to me.

Daniel Dennett, in *Consciousness Explained* (1995), defines a mystery as "a phenomenon that people don't know how to think about—yet." From the human perspective, nature's deepest mysteries have been those of origins: the origin of the universe and its physical laws, the origin of life, the origin of minds. Unifying these mysteries is the tantalizing question of how something could possibly come from nothing: how matter arises from the void, how complexity emerges from chaos, how subjective experiences derive from automata. These questions have so baffled us that for most of our history only one solution has been remotely conceivable, in which each origin is seen as the manifestation of a new ontological category. Only recently have earnest attempts been made to explain these origins within the confines of the material world without separate ontological categories for life or for minds. Although smaller mysteries still surround aspects of each origin, we are no longer entirely bewildered. The one remaining exception seems to be human consciousness, which as Dennett says, "is just about the last surviving mystery."

For this reason, it is the study of the mind/brain in which I have decided to specialize. Now, it may be said this is hardly any special-

ization at all, for this is a broad subject legitimately claimed by a number of traditional academic departments. Yet for reasons ranging from the entirely reasonable logistical demands of scientific research to occasional unintended ignorance, the theories within one department typically span only a portion of the entire spectrum of thought surrounding the subject. And since no one department has yet been able to claim the one right way to think about the subject, I feel that the interdisciplinary nature of the developing cognitive science is of utmost importance. Although it imposes some obstacles in forcing one to become fluent in the jargon and pretheoretical conceptions of each of the component disciplines, this can only strengthen the theories forged by the interdisciplinary coalition. Clearly I will be selecting a narrower focus than simply mind/brain studies for my graduate work, but I hope to maintain a high level of competence in the interdisciplinary background that surrounds my focus.

I hope to participate in a future of cognitive science that I expect will be quite exciting and full of surprises. Patricia Churchland and Ilya Farber, in "Consciousness and the Neurosciences: Philosophical and Theoretical Issues" (in *The Cognitive Neurosciences*, Michael Gazzaniga, ed., 1995), suggest that many of our current concepts about the mind are essentially prescientific, and awaiting us is a transition in which these concepts will be transformed and perhaps displaced by more precise and predictively powerful scientific concepts. Indeed, cognitive science is at a point where the right questions have yet to be asked, and a conceptual framework will evolve alongside the empirical answers that hang upon it. That science often proceeds this way is an important lesson.

Another key lesson concerns the vital relationship between science and society. Science is not done in isolation. Particularly with questions regarding the human mind, the answers will have ramifications in such sensitive areas as law, politics, ethics, and human rights, among many others. When the implications of a clearer scientific understanding of the mind are at odds with our previous conceptions, we must be willing to take an honest look at the old ideas to identify any buried motives and superstitions that underlie the conflict. We would do well to remember Bertrand Russell's thoughts (*The Autobiography of Bertrand*

Russell, 1967): "I [believe] in the value of two things: kindness and clear thinking. ...I find that much unclear thought exists as an excuse for cruelty, and that much cruelty is prompted by superstitious beliefs."

It is to promote and practice kindness and clear thinking that I intend to pursue an academic career. In this capacity, the greatest responsibility is to internalize the process and the conscience of science, and to evidence these qualities in research and teaching. I firmly believe that when performed this way, science does good, whether its benefit to society be direct, such as through cures for disease, or indirect, through the expansion of humankind's wisdom. I could not wish for more than to contribute to this cause through the study of the human mind, where a greater understanding would benefit us all on both personal and societal levels.

As I consider the path that led me from raw inspiration to a concrete culmination in this thesis, I am also reminded of the many people to whom I owe debts of gratitude. First of all, to my thesis advisor Christof Koch: for your guidance and patience and focus, for sharing in my excitement and frustrations, and for setting a superb example for how to be a scientist and for how to be a colleague. To the sources of financial support for my graduate research: a Predoctoral Fellowship from the Howard Hughes Medical Institute; the National Imagery and Mapping Agency (NIMA), now known as the National Geospatial-Intelligence Agency (NGA); the Sandia National Laboratories; the Engineering Research Centers (ERC) Program of the National Science Foundation under Award Number EEC-9402726; the National Institutes of Mental Health (NIMH); and the W.M. Keck Foundation Fund for Discovery in Basic Medical Research at Caltech. To past and present members of my thesis committee: Richard Andersen, Pietro Perona, Chris Adami, John Allman, Shin Shimojo, and those who have been my collaborators in the work presented in this thesis, Fabrizio Gabbiani and Laurent Itti. To my colleagues who have enriched my research as well as helping me to avoid work and enjoy the rest of life on occasion: Chun-Hui Mo, Jorge Jovicich (for his humor as well as for providing an alibi while I went shopping for a wedding ring), Gabriel Kreiman, McKell Carter, Sarah Farivar, Adam Hayes, Ofer Mazor, Nao Tsuchiya, Fei Fei Li, Rufin Van Rullen, Dirk Walther, Jeff Colombe, Matt Nelson, Asha Iyer. To my teachers all the way back to kindergarten whose creativity still inspires me. To ev-

eryone at Caltech who helped me keep the music going: Bill Bing, Lou Madsen, Gary Leskowitz, Matt Ashman, Clancy Rowley, Ryan Cabeen, Kjerstin Easton, Jay Bartroff, Gene Short, Lyle Chamberlain, Steve Snyder, and everyone in the Thursday Jazz Band who has come and gone through the years. To all my family, especially Mom and Dad: for the loving and nurturing environment you have always provided, and for inspiring me to strive for excellence. Finally, to my wife Kasi: for your endless love and support, and for always believing in me, there are no words, so I'll say "nothing at all."

CHAPTER 1

Introduction

1.1 We open our eyes and see

We open our eyes and we *see*.

It's so simple. Yet the human visual system achieves such computational feats that to reproduce them with the conscious "I" would confound even the most prodigious mathematical mastermind. Our brain tells us "same" for the visage of a loved one, whether that face is seen in sunlight or firelight or starlight, whether seen from left, right, near or far, whether the face has aged by a few days or a few decades; yet, show us that same face next to its brother or sister, and our brain instantly tells us "different" even when the two faces are seen in precisely the same pose and lighting. Meanwhile, our eyes make on the order of 100,000 saccadic eye movements every day, masking the oft-forgotten fact of visual life that our visual acuity decays dramatically outside the central few degrees of the visual field.

Since it is the collective conscious "I" of all of us as scientists and engineers that designs artificial machine vision systems, such systems inherit a paradox that has plagued artificial intelligence: the tasks that are simple and efficient for human observers (such as identifying that loved one's face) are monumental challenges for machine vision systems, while other tasks that are nearly impossible for human vision (differentiating a "T" from an "L" in the periphery) are trivial for a computer program. Thus, machine vision and human vision currently have complementary strengths and weaknesses. But ultimately it is critical that machine vision systems be able to assist human observers even in tasks that are tradition-

ally human-friendly yet seem computationally intractable by today's means.

Recent developments in the computational neuroscience of vision have suggested an innovative technology for addressing these goals, that of neuroscience-enabled machine vision. This thesis describes work that aims to build a more concrete understanding of the workings of specific mechanisms of the visual system, through a proof-by-example approach: we construct working computational models and test whether they match the behavior of psychophysics subjects. If (when) the models pass this test, we have gained knowledge in the neurobiology of vision, and along the way we may have built a better machine vision system as well.

Nature is fond of arranging things in *two*s, and the visual system is no exception: starting from two eyes, the world is split into two (left and right) visual half-fields, information from each being sent to two separate half-brains. Millions of years of evolution in the presence of a ubiquitous visual horizon has taught the brain to further split the world into upper and lower visual half-fields (and to pay extra attention to the lower one). From the retina, visual information also begins a divergence into two parallel processing streams, one (roughly) for motion and spatial judgments, and one for shape and color; these are variously named as the *where/what*, or *magno/parvo*, or *action/perception*, or *dorsal/ventral* pathways [Schneider, 1969, Ungerleider and Mishkin, 1982, Mishkin et al., 1983, Goodale and Milner, 1992, Milner and Goodale, 1993, Tanaka and Shimojo, 1996]. These parallel streams provide the top-level structural organization of this thesis, which contains two main parts: the first addresses visual object categorization, or *how do we know what we're looking at?*, and the second addresses visual attention *how do we know where to look?*

1.2 Categorization

One of the basic perceptual constructs of the visual system is the division of the world into discrete *objects*, separate from each other and separate from the background. Not only are the objects segregated, but they have meanings or labels attached, which can readily lead a verbal description or other overt behavior. Furthermore, the visual system often provides a hierarchy of appropriate labels, depending on the context. The *basic-level category* [Rosch et al., 1976] is what most people will answer when shown an object (or a picture of one) and asked "what is that?" This is the category level at which members share a similar shape and set of features; thus, different basic-level categories are distinguished by having dif-

ferent shapes or different constituent features. The basic-level category is a *visual* category, not a linguistic or semantic one, because categories occupying the same linguistic level are not always at the same visual level. In an oft-cited example, when people are asked to name a picture of a bird, they will likely respond with "bird," unless the picture contains an ostrich or a penguin, in which case they will be very *unlikely* to respond with "bird," but rather will use the more specific term (unless they are bird experts; see Tanaka and Taylor, 1992). This can be explained by the fact that most birds do share a common shape, while those species that are exceptions to the rule will belong to a separate basic-level category. Other examples of basic-level visual categories are human faces, four-door sedan automobiles, jogging shoes, and ball-point pens.

At levels higher than the basic level, the hierarchy of categories ultimately extends out of an exclusively visual domain and into the linguistic domain, where we find categories like animals, plants, chairs, and food. Here we find objects like *steak sandwich* and *fruit salad* that, despite sharing few visual characteristics, are both associated with the same category label "food." Objects of these superordinate-level categories can be identified and named on the basis of visual evidence, to be sure, but the category boundaries are not learned by strictly visual means.

Moving in the other direction through the hierarchy, more detailed than the basic-level category is the subordinate-level category. Examples of subordinate-level categories include male faces and female faces (subordinate categories of the basic-level category of human faces) as well as different models of four-door sedan. In addition to sharing a similar shape and set of features, members of a subordinate-level category also typically share a common spatial arrangement of those features. On the other hand, members of different subordinate-level categories within the same basic-level category are distinguished from each other by having different spatial arrangements of the same features. This is the level that is associated with expertise; in fact a common definition for visual expertise in a certain domain is the ability to perform subordinate-level tasks with the same speed and precision as basic-level tasks. Thus it is said that we are all natural-born experts at face recognition, for upon viewing a face we recall the person's name just as quickly as we identify the fact that it is a face in the first place. The same characteristic is found in trained experts in other fields, such as car enthusiasts or bird watchers. Some evidence from fMRI suggests that the same *fusiform face area* [Kanwisher et al., 1997] that is involved in face processing may also be involved in expert processing of subordinate-level category information [Gauthier et al., 1997,

1999, 2000b,a, Tarr and Gauthier, 2000].

Part I of this thesis focuses on computational models of subordinate-level categorization. We begin in Chapter 2 by introducing several simple sets of schematic, line-drawn objects, each sharing a common set of features and differing in the spatial arrangements of those features. Subsequent chapters explore the ways in which such stimuli might be processed by the visual system: how the raw sensory input from the retina is transformed into a compact intermediate representation (Chapter 5), what the nature is of this intermediate representation (Chapter 4), and finally how this representation can be used to reach a categorization decision about the input stimulus (Chapter 3).

1.3 Attention

A neuron anywhere in the visual system undoubtedly exemplifies another of nature's dichotomies: bottom-up and top-down influences. Bottom-up processes are typically thought of as stimulus-driven, unconscious, automatic, not subject to voluntary control, and produced by anatomical feed-forward connections. Examples of bottom-up processes in vision include the detection of a flash of light, the "pop-out" of a red object amongst a green background, or the identification of a face. Jerry Fodor [1985] has used the term *cognitively impenetrable* to describe early bottom-up processes: we cannot voluntarily alter them, and furthermore we cannot determine the mechanisms for their action by introspection. Although bottom-up visual processing seems subjectively simple because it requires no voluntary effort and is cognitively impenetrable in any case, any computational neuroscientist or machine vision engineer will attest that the underlying mechanisms are far from simple. In a metaphor from computer programming, this is classic information hiding: the visual system shields the conscious "I" from a mass of implementation complexity with a trivially simple interface (*i.e.*, we open our eyes and *see*).

Standing in contrast to bottom-up processes are top-down processes: driven by cognitive beliefs, conscious, subject to voluntary control, produced by anatomical feedback connections. An example of a top-down process is the situation where one views an ambiguous scene, with no identifiable subject, yet when one is cued to look for a particular object (such as the well-known dalmatian dog in the halftone image), the previously unseen object now becomes readily apparent. Moreover, upon viewing the same scene again at a later time, the object is perceived

without delay. Another common example of top-down processing is the Rorschach inkblot test, in which people are shown inkblots that have no "true" interpretation, yet high-level cognitive states force an interpretation of the image from the top down.

Although bottom-up and top-down processing are eternally entwined, a computationally tractable modeling effort begins by treating the two separately at first. Thus Part II is aimed at the continued development a computational model of bottom-up component of visual attention, which we compared with the locations fixated by human observers while they freely viewed several types of images. Chapter 6 shows that this model can account for a significant fraction of the fixation locations, even in apparently high-level tasks, like looking for interesting things in an image. Chapters 7 and 8 introduce two new model components that improved the model's realism in replicating biological information processing, with a focus on nonlinear interactions within and across the neural representations of basic visual cues (like horizontal and vertical orientations). Interestingly, this improved biological realism increased the models ability to predict locations either looked at or pointed to by human observers with very high statistical significance. Finally, Chapter 9 introduces a "mouse-clicking" method as a lightweight alternative to eye tracking.

Part I

Visual object categorization

CHAPTER 2

Categorization psychophysics with parametric stimuli

2.1 Introduction

Visual object recognition and categorization are critically important to the survival of many animal species, notably humans. These processes constitute an impressive computational feat, in that we are able to, on the one hand, lump together as "same" a set of views of an object seen under different conditions that produce radically different patterns of light on the retina, yet on the other hand, mark as "different" two views of different objects under similar viewing conditions that produce very similar patterns of light falling on the retina.

In the last thirty years, research in mathematical psychology has discovered much about the processes of visual categorization [*e.g.*, Reed, 1972, Nosofsky, 1984, 1991, Ashby, 1992a, Ashby and Maddox, 1993, Smith and Minda, 1998, Ashby and Waldron, 1999] by combining the techniques of visual psychophysics and computational modeling to develop high-level theories of categorization. Despite the predictive success of these theories, there exists a gap between the descriptive framework of the models, and our current knowledge of the neuronal mechanisms involved in categorization. An important aim therefore is to shorten this gap by extending models so that their implementations are reasonable in light of recent developments in the neurophysiology of object recognition and categorization [Kanwisher, McDermott, and Chun, 1997, Ishai, Ungerleider, Martin, Schouten, and

Haxby, 1999, Freedman, Riesenhuber, Poggio, and Miller, 2001, Sigala and Logothetis, 2002, Op de Beeck, Wagemans, and Vogels, 2001, Ashby and Ell, 2001].

Most categorization models assume, perhaps tacitly, a categorization process in which

- the immediate sensory representations of incoming stimuli occupy a very high-dimensional space (for vision, this comes to millions of dimensions when we consider that output of each retinal ganglion cell amounts to a single dimension);

- this very high-dimensional representation is transformed into an intermediate space of lower dimensionality by combining the simple features of the early representation (such as oriented edges) into more complex features (such as *T*- or *L*-junctions or simple shapes);

- finally, a computational process operates in this lower-dimensional space and produces an explicit categorization result that can be the basis for a behavioral response (such as a button-press in a psychophysics experiment).

Here, we address each of these three key steps with categorization models informed by psychophysics experiments and neurobiology. We will visit the topics in reverse order: our goal is to describe how a categorization decision is made for a visual object, and each of the next three chapters takes a step toward filling in the details of this process back to the retinal input. First, in Chapter 3 we compare several existing models of visual object categorization and introduce a new *roaming exemplar model* (RXM) that highlights the importance of certain model characteristics in matching human behavior. Second, in Chapter 4 we delve into the nature of the intermediate representation using *multidimensional scaling* (MDS). Third and last, many models in the psychological literature ignore the transformation which turns raw sensory data into a succinct set of nameable features; we explore this step in detail in Chapter 5 using a model of early vision similar to "HMAX" [Riesenhuber and Poggio, 1999] to drive subsequent phases of the categorization process.

Eighteen psychophysics subjects (ages 18–25) from the Caltech community participated as paid volunteers in the experiments described in the following three chapters. Informed consent was obtained from all subjects, and experimental procedures were approved by the California Institute of Technology's Committee for the Protection of Human Subjects.

2.2 Psychophysical methods

Like many other studies of visual object categorization, we used schematic line-drawn stimuli. Many insights into the principles of categorization have been obtained with minimalistic stimuli, such as the circles shown in Figure 2.1. We used three types of schematic, line-drawn visual stimuli with somewhat more complexity (Figure 2.2): Brunswik faces and tropical fish outlines, which have been used previously, plus a new set of "cartoon face" images. Each type of visual object was parameterized along four dimensions comprising the *stimulus parameter space*. For each object type, different sets of objects were assigned to *configurations*, which contained equal numbers of *training exemplars* assigned to each of two categories, as well as an additional number of *test exemplars*. The training exemplars from the two categories were always chosen so as to be linearly separable in the objects' parameter space; that is, the members of the two categories could be separated by some 3-D hyperplane in the 4-D parameter space.

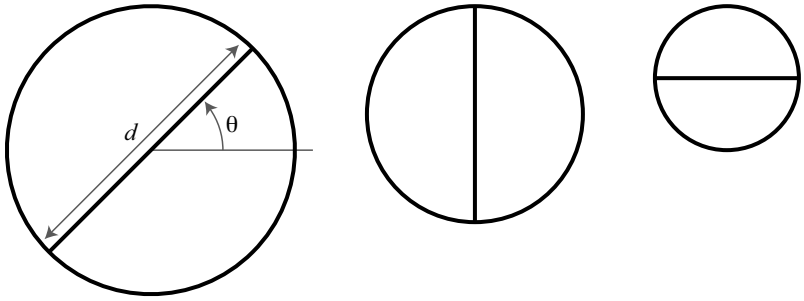


Figure 2.1. An example of the type of parametric stimuli used in previous studies [*e.g.*, Maddox and Ashby, 1993] of visual object categorization. In this example, the objects are circles with a bisecting line defined by two features: the diameter d of the circle and the angle θ of the bisecting line. Experiments using these stimuli typically involve categories devised so that both features must be analyzed in order to determine category membership. For example, category 1 might include all small circles as well as some medium-sized circles with nearly-horizontal bisectors, while category 2 would include all large circles as well as medium-sized circles with nearly-vertical bisectors. In such an experiment, the question of interest would be to quantitatively understand how observers categorize the ambiguous objects of medium size with diagonal bisectors.

2.2.1 Brunswik faces

These simple line-drawn face stimuli (Figure 2.2a; Brunswik and Reiter, 1937) have been used frequently in categorization experiments both with human [Reed, 1972, Nosofsky, 1991] and non-human observers (pigeons, Huber and Lenz, 1996; mon-

keys, Sigala et al., 2002). Each face consists of a simple ovaloid outline with internal features defined by (compressed) circles and straight lines. The faces are parameterized by *eye height* (EH; the vertical distance from the centers of the eyes to the center of the face), *eye separation* (ES; the horizontal distance separating the centers of the eyes), *nose length* (NL; the vertical length of the nose line), and *mouth height* (MH; the vertical distance from the center of the face to the mouth line).

2.2.2 Cartoon faces

These stimuli (Figure 2.2b) were introduced in an fMRI study [Jovicich, Peters, Koch, Chang, and Ernst, 2000] that showed them to produce stronger activation in the human fusiform face area [Kanwisher et al., 1997] than did Brunswik faces (see Section 2.3 below). The cartoon faces extend the Brunswik faces in several ways to make the faces appear more human: a simple band of hair is added around the top of the head, the size and dilation of the pupils may be varied, eyebrows are added above the eyes, the nose outline is defined by an extended open contour, and the mouth is defined as a Bezier curve rather than a straight line. To control these additional features, the cartoon faces have a total of 28 stimulus parameters; however, in the present study only the four parameters corresponding to the Brunswik face dimensions were varied, while the other 24 parameters were held constant.

2.2.3 Tropical fish outlines

These line-drawn images (Figure 2.2c) were first used to offer a completely novel stimulus set to monkey observers in a categorization task [Sigala et al., 2002]. Other fish images have been used previously in studies of categorization in people and pigeons [Hernstein and de Villiers, 1980] and in monkeys [Vogels, 1999]. Each fish image is composed of four cubic spline curves that were fitted to scanned outlines of tropical fish. By adjusting one control point of each of the curves, four features of the outlines could be smoothly deformed: the dorsal fin (DF), tail fin (TF), ventral fin (VF), and mouth area (MA).

2.2.4 Stimulus rendering

All of the stimuli described here were generated and displayed to subjects using `GroovX`, a custom software software designed for psychophysics and object-oriented graphics. The package is licensed under the GNU General Public Li-

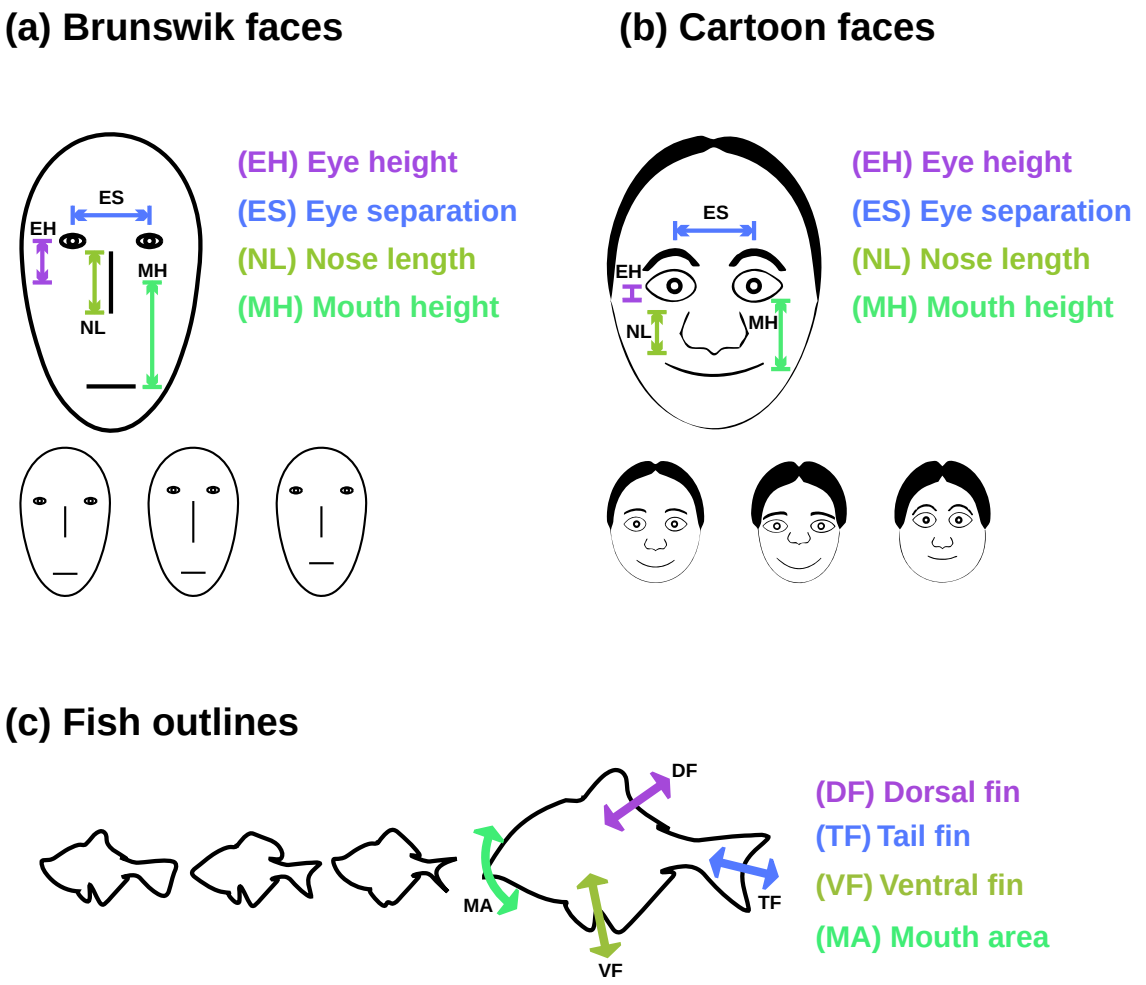


Figure 2.2. Three object types, each with four stimulus parameters controlling that object type, were used in similarity and categorization psychophysics tasks. Three sample objects of each type demonstrate the typical ranges of the parameters. **(a)** *Brunswik faces*. **(b)** *Cartoon faces*. Although these faces are described by 28 parameters, the present study used only the 4 parameters corresponding to those in (a). **(c)** *Fish outlines*.

cense (GPL), and is freely available for download from the internet. Currently documentation and source code may be found online at this web address: `http://www.klab.caltech.edu/rjpeters/groovx/`. In the event that this software later moved to a different location it should be locatable via a number of popular internet search engines using the keyword `GroovX`.

2.3 Neural representation of schematic stimuli

To provide a foundation for our categorization model results to be eventually generalized into the domain of natural (*i.e.*, photorealistic rather than schematic or

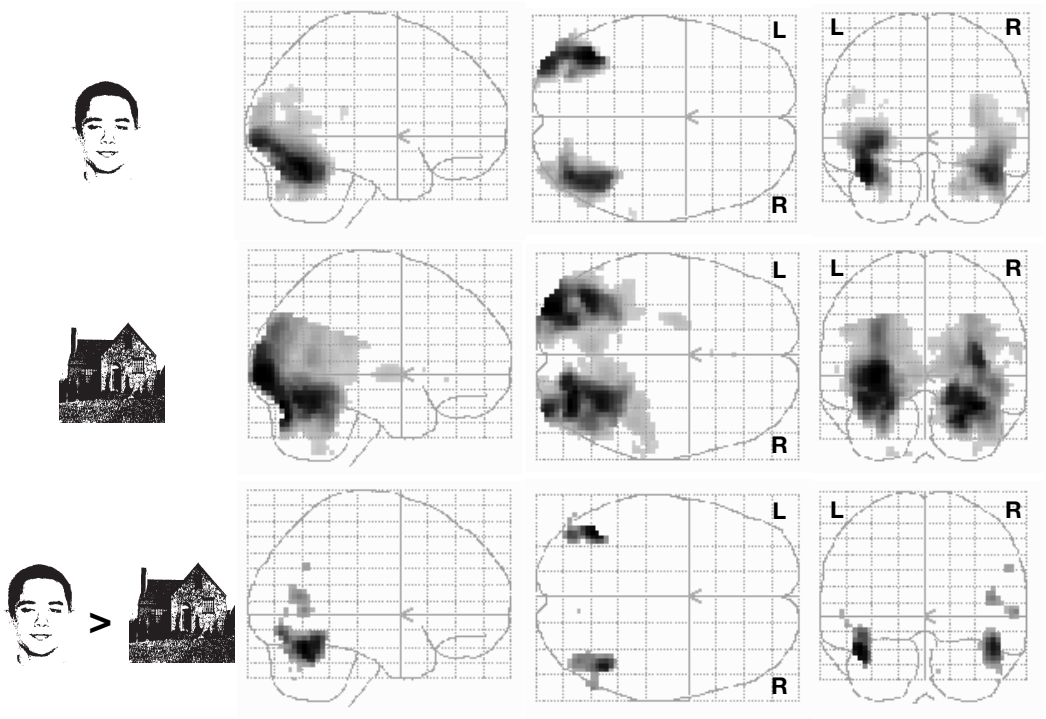


Figure 2.3. BOLD activity observed with fMRI while subjects viewed photos of either faces (top row) or houses (middle row), and the contrast between two conditions showing regions which were more active during face viewing (bottom row). This contrast reveals bilateral activation in the *lateral occipital cortex* (LO) and *fusiform face area* (FFA).

line-drawn) visual stimuli, we used an fMRI experiment [Jovicich et al., 2000] to assess whether the schematic Brunswik and cartoon faces produced neural activation in the same *fusiform face area* (FFA) that is known to be activated by photos of real faces [Kanwisher et al., 1997]. Four healthy adult subjects (2 females, 2 males, ages 18–23) participated; all were right-handed and had corrected-to-normal vision. The stimulus sequences contained six 30 s stimulus epochs interleaved with seven 20 s fixation epochs. In each stimulus epoch, subjects passively viewed rapid sequences of either faces or houses (2.5 Hz presentation rate, 75 images per epoch).

Whole brain single shot T2*-weighted spiral functional images were acquired using the manufacturer's head birdcage coil on a 1.5-Tesla scanner (General Electric Signa, Milwaukee, WI). Imaging parameters were: TE=50 ms, TR=2500 ms, 3.125×3.125 mm in-plane resolution, 4 mm-thick axial slices, 1 mm slice gap. Data were processed using standard procedures in SPM99b. Before statistical analysis, data were motion corrected, Tailarach normalized, and spatially smoothed. The data were analyzed using a fixed-effects statistical model comprising subject-specific effects (signal change during face viewing and signal change during house

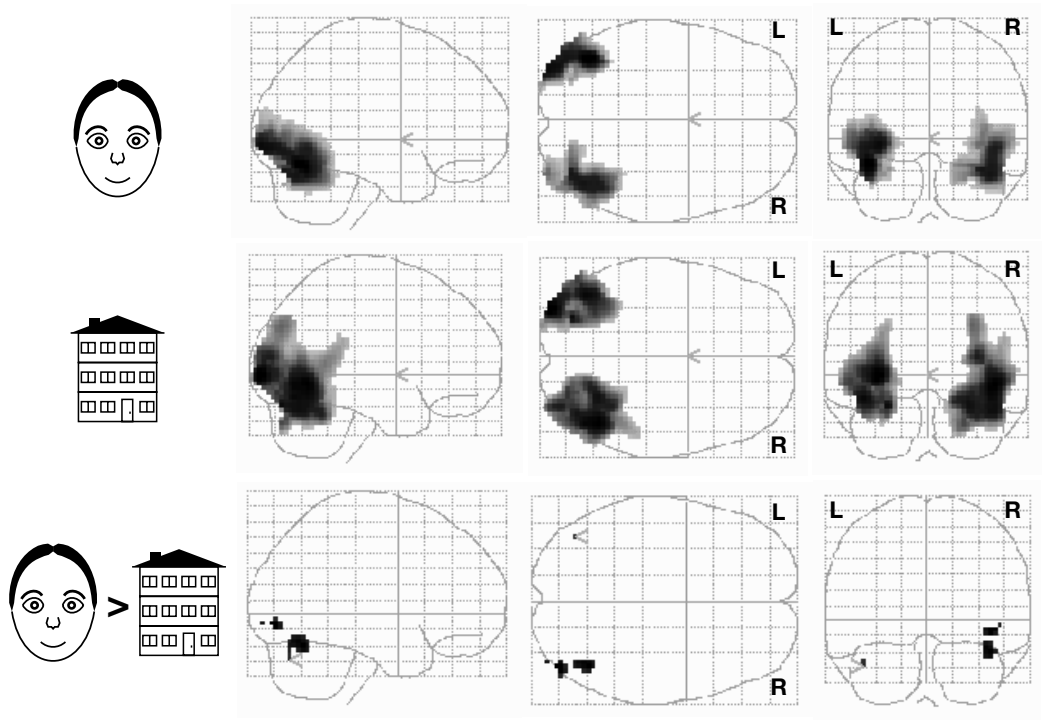


Figure 2.4. Same conditions as in Figure 2.3, except that cartoon faces and cartoon house were used in place of the photos. Compared to the condition with face and house photos, activation here in LO and FFA is reduced on the right side, and is nearly absent on the left side.

viewing). Activation was defined as the difference in the signal changes between the face and house conditions. For each subject, data were scaled to the global brain mean and analyzed separately to locate the areas which responded more to faces than to houses during the different stimuli sequences. The strength of the effect was then averaged in these areas across subjects for each condition.

When we looked for areas that were significantly more activated by photos of real faces than by photos of real houses (Figure 2.3)[1], we found strong bilateral activation in the lateral occipital (LO) cortex ($p < 0.001$, uncorrected) as well as in the inferior temporal gyrus consistent with the FFA ($p < 0.001$, uncorrected). For comparison with the Brunswik and cartoon faces, we designed schematic house stimuli with different levels of complexity to match the subjective complexity of the different face stimuli. In the analogous face/house comparisons, we found FFA activation for the cartoon faces that was strongly right-side dominant (Figure 2.4), and we found no detectable FFA activation for the Brunswik faces (Figure 2.5). This does not rule out the possibility that Brunswik faces activate neurons in the

[1]Thanks to Nancy Kanwisher for the use of the face and house photo databases.

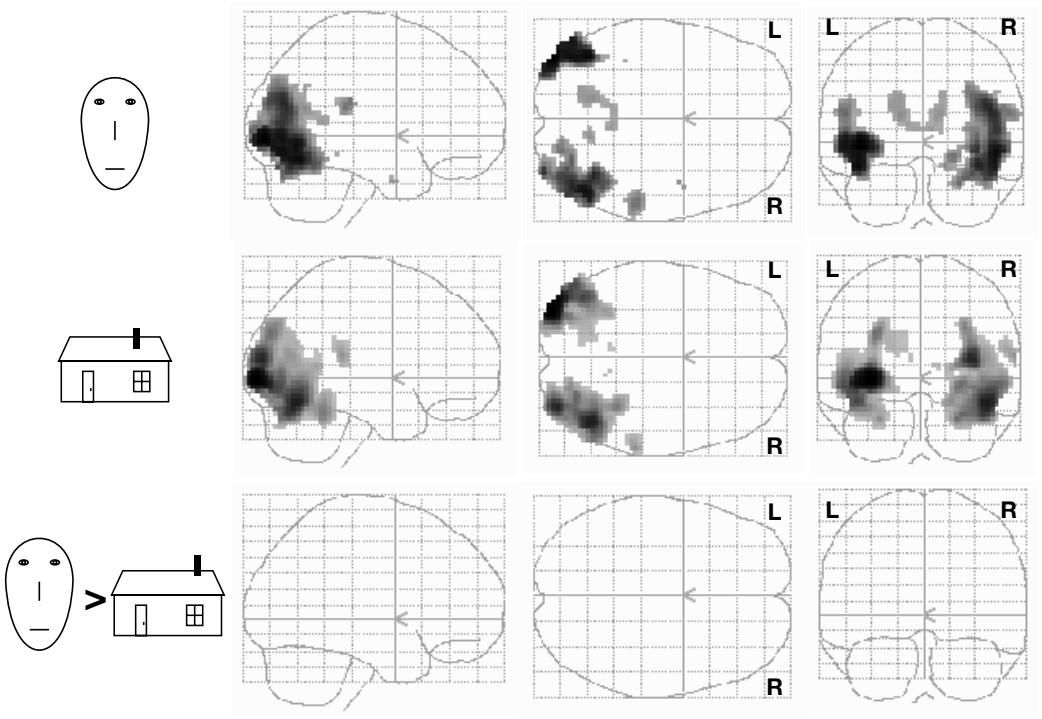


Figure 2.5. Same conditions as in Figures 2.3 and 2.4, except that Brunswik faces and simple house cartoons were used. No activation was observed in either LO or FFA.

FFA; since there is less variability between different Brunswik faces than between different cartoon faces or face photos, it would be expected that the Brunswik faces would activate a smaller pool of FFA neurons, and the lack of observable activity might simply indicate that these neurons are undergoing adaptation during the course of each 30 s epoch. In any case, our results do offer positive confirmation that at least the cartoon faces activate neurons within the same FFA pool as those activated by real face photos.

CHAPTER 3

Categorization models

3.1 Introduction

This chapter describes a number of models that attempt to mimic human categorization decisions using a mechanism based on the multidimensional representation of incoming stimuli, plus possible auxiliary representations, such as memory traces. This process is typically controlled by a number of free parameters, which are fitted with the goal of matching human categorization behavior. However, a simple statistical comparison between models—even after accounting for the number of free parameters—may ignore important differences in the neurobiological implications of the models. For example, one successful model, the *generalized context model* (GCM; Nosofsky, 1984), assumes that all training images are stored in memory; a literal interpretation of the GCM might conclude that the neuronal substrate of categorization also scales linearly with the number of exemplars in a category, or that categorization in biological systems involves only simple memorization, without any category-level abstraction [Knowlton, 1999]. To provide a more detailed look at such issues, we introduce a *roaming exemplar model* (RXM) that draws from neural networks [Poggio and Girosi, 1990, Rosseel, 1996] and exemplar-based models of categorization [Nosofsky, 1991, Kruschke, 1992, Nosofsky, Kruschke, and McKinley, 1992]. The RXM also has much in common with the *striatal pattern classifier* (SPC) of Ashby and Waldron [1999], including the fact that its memory traces are free parameters. This stands in contrast to previous exemplar-based models, and hence neurobiological plausibility can be

assessed directly by accounting for numbers of free parameters when comparing fitted models.

3.2 Categorization psychophysics tasks

Subjects participated in categorization experiments consisting of a training phase and a testing phase. In both phases, subjects viewed a series of objects presented one at a time. Each object was shown for 2 s, followed by 2 s of blank screen. During each 4 s trial, subjects pressed one of two buttons indicating to which category the object belonged. In the training phase, subjects were shown only the two categories' training exemplars, and were given feedback in the form of a high- or low-pitch tone indicating whether their response was correct or incorrect, respectively. Subjects performed training blocks of 100 trials until they scored $\geq$ 85% correct on a single block. Next, they moved into the testing phase, in which they were shown the previously unseen test exemplars in addition to the training exemplars that they had viewed during the training phase. Subjects received no feedback on their responses during the testing phase.

In Experiment 1, the values for each stimulus dimension were quantized to three possible values for each dimension, so that the set of possible objects lay on a $3\times3\times3\times3$ grid in stimulus parameter space. The configuration of 20 objects on this grid (Figure 3.1a) followed that used in Nosofsky [1991] and Sigala et al. [2002], with five training exemplars for each category, plus ten test exemplars that included the two category prototypes. For each set of objects, each of the four stimulus parameters for that object type was assigned to one of the four generic dimensions in the stimulus configuration shown in Figure 3.1a. It is significant how the parameters are assigned, since each generic dimension carries different information about category membership. For example, the categories were linearly separable in projections onto 2-D planes for pairs of stimulus dimensions $(1,2)$, $(1,3)$, and $(1,4)$, so dimension 1 was more informative about an object's category than were the other dimensions. In all, five sets of stimuli were used in Experiment 1. These included three sets of Brunswik faces in which the stimulus parameters were assigned to the generic dimensions in different orderings ({EH, ES, NL, MH}, {NL, MH, EH, ES}, and {MH, EH, NL, ES}), a set of cartoon faces ({EH, ES, NL, ML}), and a set of fish outlines ({TF, VF, DF, MA}). The twenty objects in each of these configurations are shown in Figures 3.2, 3.3, 3.4, 3.5, and 3.6.

In Experiment 2, a larger configuration of 80 objects was used (Figure 3.7), with

10 training exemplars for each of the two categories, plus 60 test exemplars. The exemplars were arranged on a 7×7×7×7 grid in the stimulus parameter space. There were 12 such sets, identical except that the discretization grid of each set was rotated through different angles ($\theta = n \cdot 15°, n \in [0 \ldots 11]$) in the eye-height/eye-separation plane of parameter space.

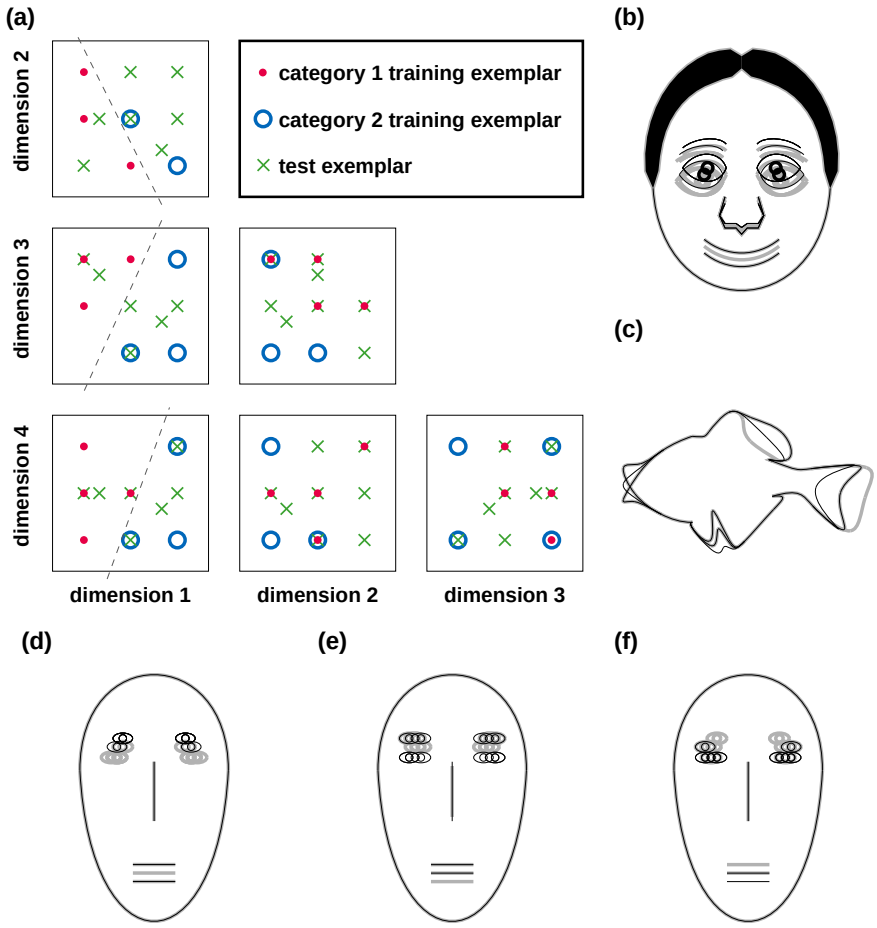


Figure 3.1. Experiment 1 used five 20-object sets, each defined in a 4-D parameter space. **(a)** The abstract configuration is shown in projections onto the six possible pairs of dimensions. All exemplars fall on a 3×3×3×3 grid, except for the two category prototypes, which were among the test exemplars. Dashed lines indicate where the two categories' training exemplars are linearly separable. **(b-f)** For illustration, the training exemplars of category one (thin black lines) are superimposed upon those of category two (thick gray lines). **(b)** Cartoon faces with dimensions {1=EH, 2=ES, 3=NL, 4=ML} (see also Figure 3.2). **(c)** Fish outlines {TF, VF, DF, MA} (Figure 3.3). **(d)** Brunswik faces {EH, ES, NL, MH} (Figure 3.4). **(e)** Brunswik faces {NL, MH, EH, ES} (Figure 3.5), and **(f)** Brunswik faces {MH, EH, NL, ES} (Figure 3.6). See Figure 2.2 for abbreviations.

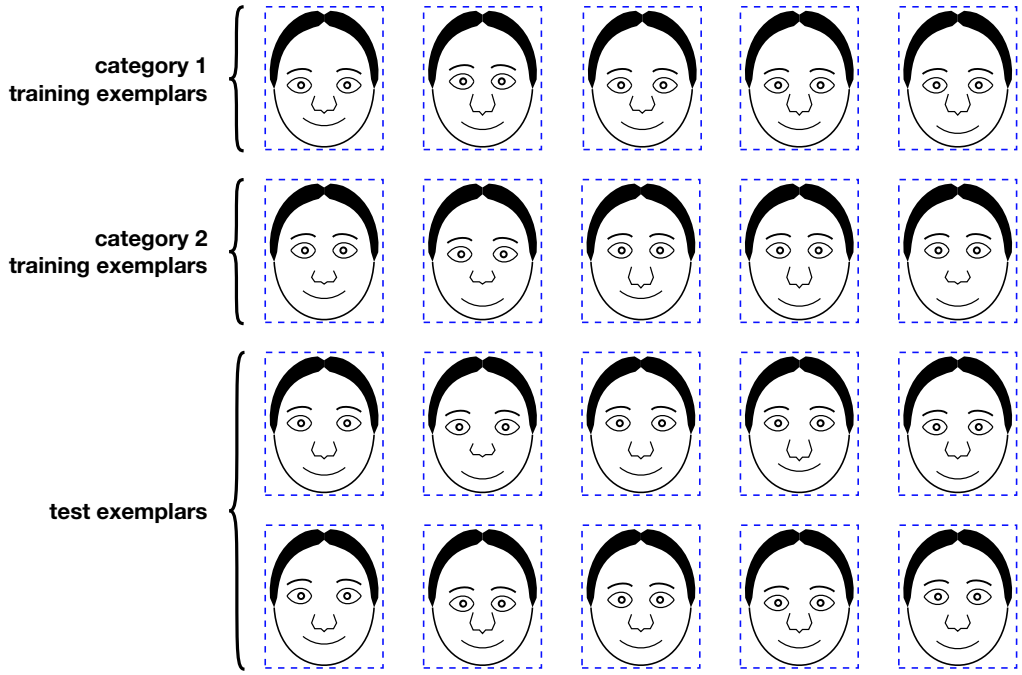


Figure 3.2. Cartoon faces {EH, ES, NL, ML} used in experiment 1.

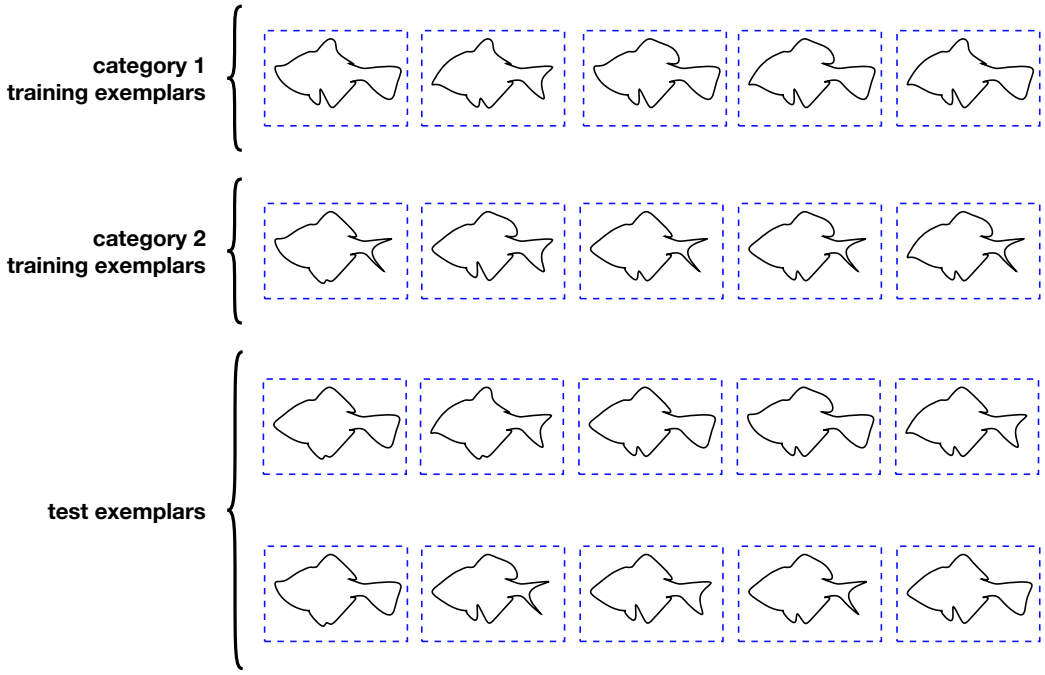


Figure 3.3. Fish outlines {TF, VF, DF, MA} used in experiment 1.

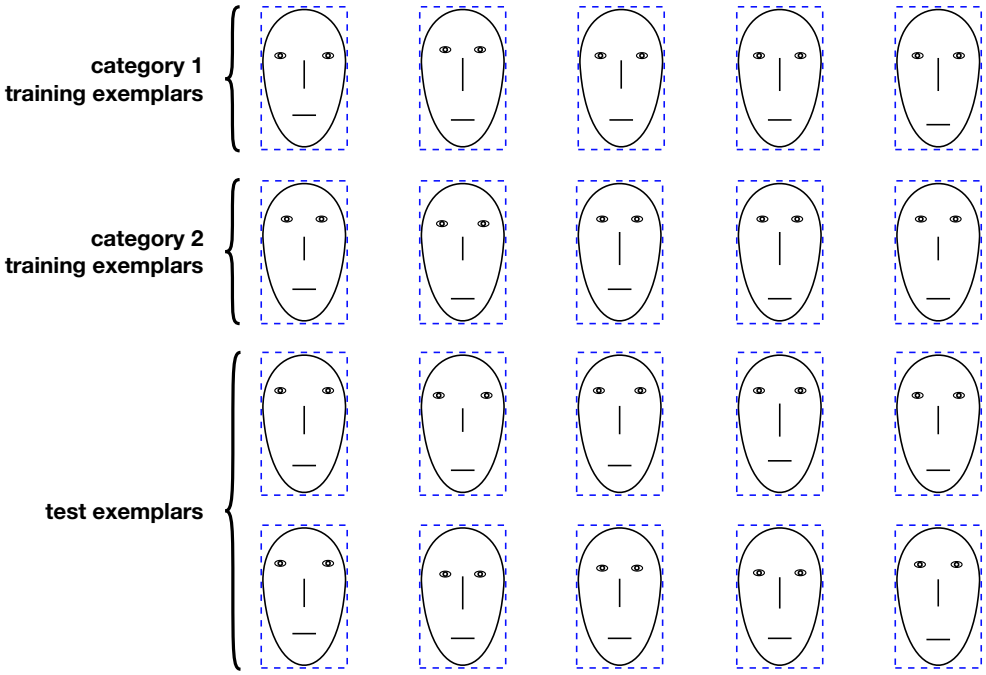


Figure 3.4. Brunswik faces {EH, ES, NL, MH} used in experiment 1.

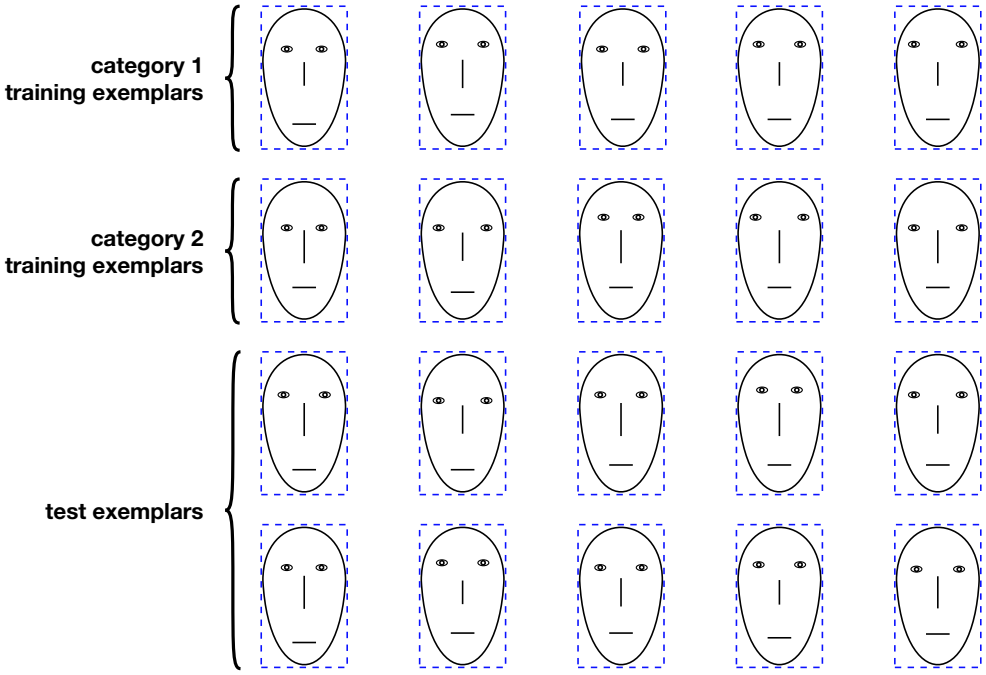


Figure 3.5. Brunswik faces {NL, MH, EH, ES} used in experiment 1.

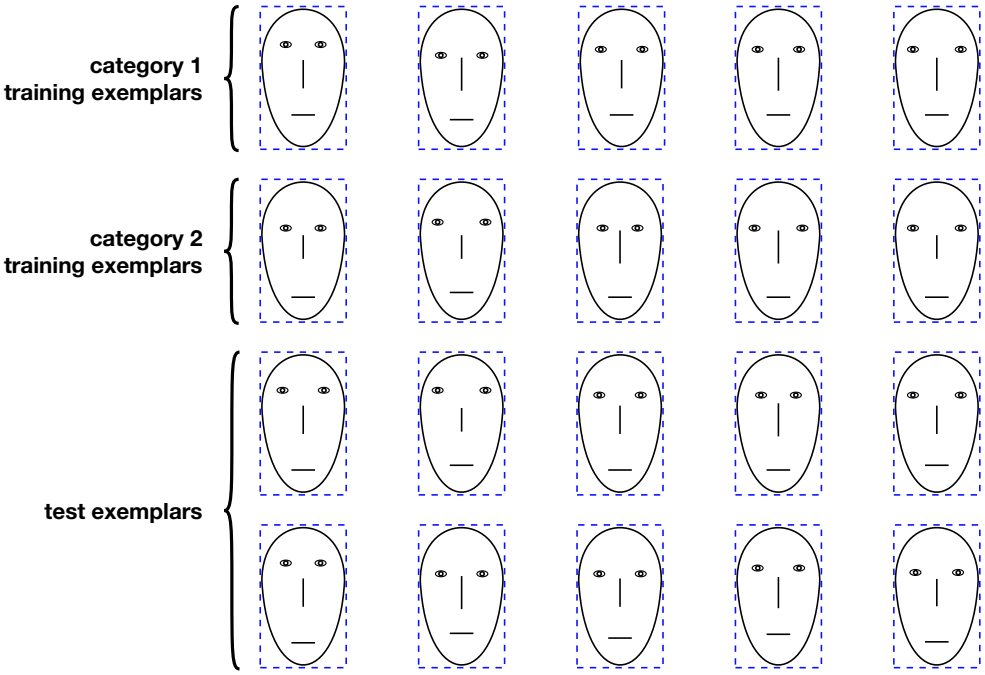


Figure 3.6. Brunswik faces {MH, EH, NL, ES} used in experiment 1.

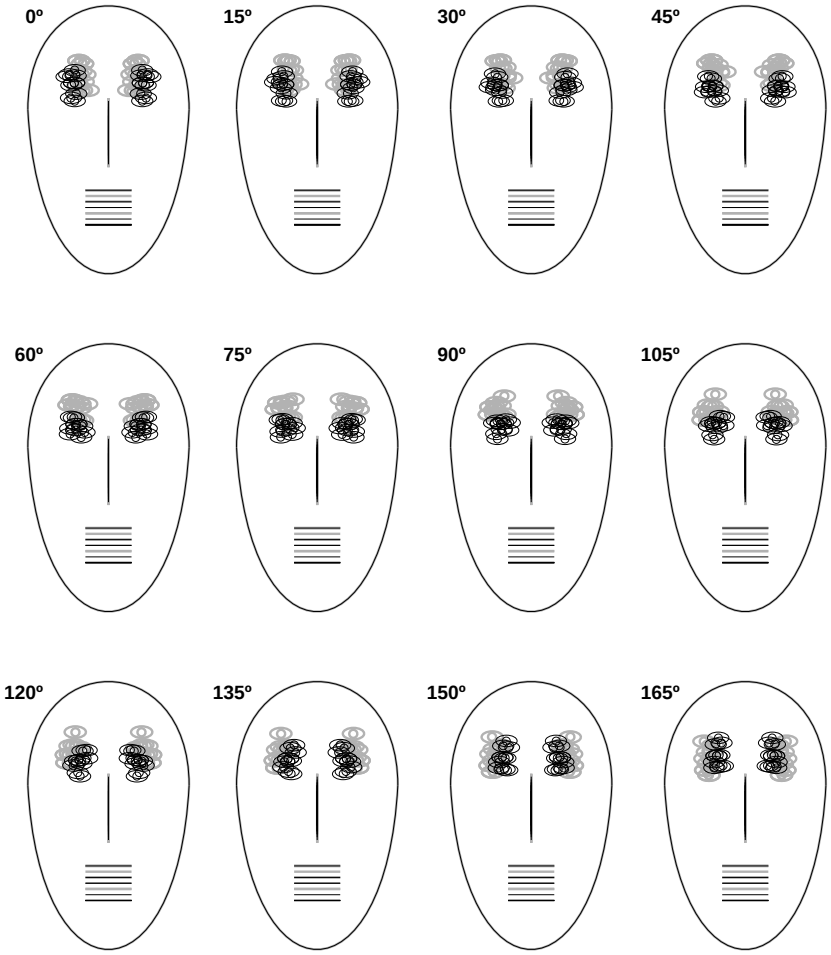


Figure 3.7. Experiment 2 used these 12 sets of Brunswik faces. Each image shows the 10 training exemplars of category one (thin black lines) superimposed upon the 10 training exemplars of category two (thick gray lines). The sets differ only in the angle by which the objects are rotated in eye height-eye separation plane of feature space.

3.3 Categorization models

We tested several categorization models by fitting them to match the human observers' response profiles from the testing phase of the categorization tasks. Each model receives input in a 4-D feature space (*i.e.*, not image space), and produces an output that represents a categorization probability for the input object. The models we tested fall into several categories, each of which proposes a unique architecture for the categorization process (see Figure 3.8), with different free parameters, and different assumptions about the memory usage of the system being modeled. These factors must be weighed along with the raw goodness-of-fit when assessing

the neurobiological plausibility of the different models.

In general, the categorization models assume the following:

- that each exemplar x has a unique representation in an R-dimensional space [Ashby, 1992b],

  $$\boldsymbol{x} = (x_1, \ldots, x_R),$$

  whose components may be drawn either from the original stimulus configuration, from a multidimensional scaling configuration (see Chapter 4), or from a configuration based on features extracted from an early-vision model (chapter 5), and

- that each category is defined by N training exemplars

  $$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}.$$

3.3.1 Exemplar models

Exemplar models associate memory traces of M (with $1 \leq M \leq N$) stored exemplars[1]

$$\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M\}$$

with each category. Several model subtypes differ in the way that these stored exemplars are selected:

- *All-exemplar models* (Figure 3.8a) assume $M = N$, and $\boldsymbol{y}_i = \boldsymbol{x}_i$. All of the training exemplars are explicitly stored in memory, so these models have a high memory demand that is linear in the number of training exemplars. All-exemplar models include the *average-distance model* (ADM, Reed, 1972) and *generalized context model* (GCM; Nosofsky, 1991).

- *Prototype (one-exemplar) models* (Figure 3.8b) assume $M = 1$; each category stores only the arithmetic mean of the category's training exemplars, $\boldsymbol{y}_1 = \frac{1}{N}\sum_i \boldsymbol{x}_i$. These models have low and constant memory demand, independent of the number of training exemplars; however, the models imply a more complex computational mechanism to estimate the prototype during trial-by-trial

[1] Our usage of the term "exemplar" to denote stored memory traces reflects a meaning of *ideal meaning or pattern* or *prototype*, rather than a strict meaning of *previous seen stimulus*. For example, in the RXM, the stored exemplars are generalizations of the memory traces used in all-exemplar or prototype models, and are most likely not previously seen stimuli.

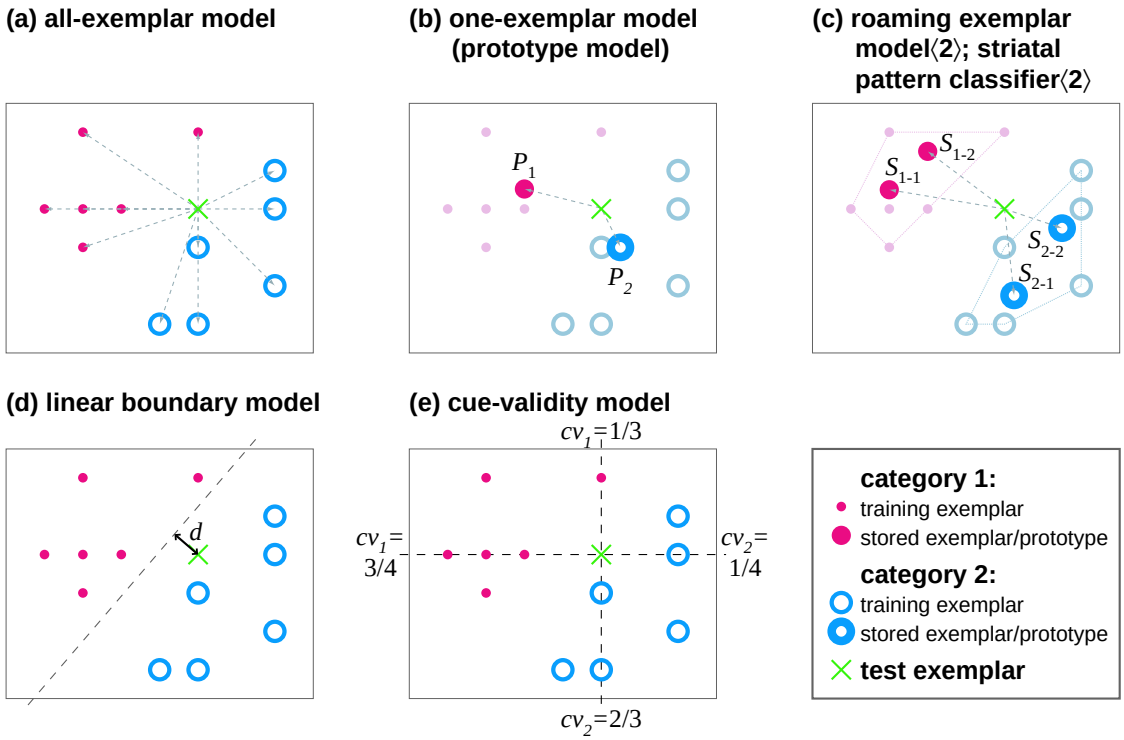


Figure 3.8. Schematic depictions of several kinds of categorization models. Each diagram shows a hypothetical set of training exemplars from two categories (• and ○) in a 2-D feature space, plus a test exemplar (×) which is to be classified. **(a,b,c)** Three types of models which rely on distances (indicated by dashed lines) between a test exemplar and each stored exemplar from both categories: **(a)** *all-exemplar model*, in which the set of stored exemplars is just the set of training exemplars; **(b)** *one-exemplar*, or *prototype model*, in which the single stored exemplar per category is the arithmetic mean of that category's training exemplars; **(c)** *roaming-exemplar model*$\langle M \rangle$ (RXM$\langle M \rangle$) and *striatal pattern classifier*$\langle M \rangle$ (SPC$\langle M \rangle$), in which each category has M (in this case, $M = 2$) stored exemplars, which must lie within the polygon that circumscribes the training exemplars (dotted lines). The RXM$\langle M \rangle$ uses a summed-similarity decision rule, while the SPC$\langle M \rangle$ uses a uses a nearest-neighbor decision rule. **(d)** *Linear boundary model*, in which the model uses a linear boundary that separates the categories to classify test exemplars according to the side of this boundary on which they fall. **(e)** *Cue-validity model*, which classifies a test exemplar according to the total cue-validity across all features; the cue-validity cv_i for category i of a given feature is the posterior probability of an exemplar with that feature belonging to category i (values of cv_1 and cv_2 are shown).

exposure to the training exemplars. Prototype models include the *weighted prototype model* (WPM; Reed, 1972) and the *weighted prototype similarity model* (WPSM; Nosofsky, 1991).

- In the proposed *roaming-exemplar model*$\langle M \rangle$ (RXM$\langle M \rangle$, Figure 3.8c), each category stores M exemplars, each of which is a linear combination of the training exemplars for that category, $\boldsymbol{y}_j = \sum_i w_{ij} \boldsymbol{x}_i$. Under the neurobiological consideration that neurons do not represent objects far different from those that have been previously observed, the stored exemplars are restricted to a region circumscribed by the training exemplars, so the weights are constrained by $w_{ij} \geq 0$ and $\sum_i w_{ij} = 1$ for all j. The number of stored exemplars M is *not* a free parameter of a given RXM$\langle M \rangle$, but the stimulus parameters of those stored exemplars *are* free parameters of the model. Thus, when the RXM is fitted to a dataset, the number of stored exemplars is chosen and fixed at the start, although RXM$\langle M \rangle$'s with different (fixed) values of M may be fitted to the same dataset. The memory demand of the RXM$\langle M \rangle$ varies between that of the prototype models (for $M = 1$) and that of the all-exemplar models (for $M = N$); the computational complexity is similar to that of the prototype models, since some mechanism must adjust the stored exemplars during training.

Next, the exemplar model computes a similarity measure between the test exemplar $\boldsymbol{x}$ and each of the stored exemplars $\boldsymbol{y}$, based on a weighted Euclidean distance:

$$d_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_j \alpha_j (x_j - y_j)^2},$$

with $\alpha_j \geq 0$ and $\sum \alpha_j = 1$ (other metrics are possible; *e.g.*, Ashby and Maddox 1993). The coefficients α_j, called *attentional weights*, are intended to model the ability of human observers to attend preferentially to the most task-relevant stimulus features. The similarity s decays with the distance d, either linearly ($s = -d$, as in the RXM, ADM and WPM), or exponentially ($s = e^{-cd}$, as in the GCM and WPSM; see Shepard, 1987).

Then, for each test exemplar $\boldsymbol{x}$, the evidence E_i for category C_i is given as the sum of similarities between $\boldsymbol{x}$ and the M stored exemplars $\boldsymbol{y}_j^i$ of that category:

$$E_i(\boldsymbol{x}) = \sum_{j=1}^{M} s(\boldsymbol{x}, \boldsymbol{y}_j^i).$$

Finally, the model's categorization of $\boldsymbol{x}$ is based on the expression

$$E_1(\boldsymbol{x}) - E_2(\boldsymbol{x}) + n > t,$$

where n represents zero-mean Gaussian noise with variance σ^2, and t is a threshold parameter; $\boldsymbol{x}$ is assigned to category 1 if this expression is true, otherwise to category 2.

The free parameters of the exemplar models are thus $(\boldsymbol{\alpha}, c, t, \sigma)$, plus $2M$ stored exemplars for the RXM$\langle M \rangle$.

3.3.2 Striatal pattern classifier

The RXM shares a very similar mathematical formulation with the *striatal pattern classifier* (SPC) proposed by Ashby and Waldron [1999], although the mathematical elements have been treated with different neurobiological interpretations [Ashby and Ell, 2001]. Both kinds of model rely on a set of units that represent different locations in feature space, but the models differ in how each category's evidence is computed for a given test exemplar. The exemplar models compute the sum of similarities between the test exemplar and each stored exemplar, whereas the SPC associates a test exemplar with the category of the nearest striatal pattern (in this respect the SPC resembles a k-nearest neighbor model with $k = 1$). Both the SPC and the RXM use a similarity measure that decays linearly with distance. In order to maintain a formal similarity with the other models, we used the following decision rule for the SPC: for each test exemplar, the evidence for each category is given by the maximum of the similarities between the test exemplar and that category's stored exemplars. Thus, in the case of one stored exemplar per category, the SPC$\langle 1 \rangle$ and the RXM$\langle 1 \rangle$ form identical decision surfaces. However, with $M > 1$, the SPC$\langle M \rangle$ has a piecewise-linear boundary, while the RXM$\langle M \rangle$ has a curved decision boundary.

3.3.3 Boundary models

Decision bound theory [Ashby and Maddox, 1993] proposes that human perceptions of category exemplars are instances of random variables with multivariate normal distributions. Given a particular perception, the optimal decision strategy is to choose the category of which that perception was more likely an instance. Thus the decision boundary (the locus where both categories have equal proba-

bility densities) falls along the intersection of the graphs of the two probability density surfaces. If the covariance matrices of the exemplar distributions are identical for the two categories, then the decision boundary is a linear surface (*i.e.*, a hyperplane); otherwise, it is a quadratic surface.

We tested the probit linear model (PBI; Figure 3.8d; Ashby and Gott, 1988), which is trained to separate the categories' training exemplars with a boundary described by a normal vector $\boldsymbol{b}$ and a threshold t. Following training, a test exemplar $\boldsymbol{x}$ is classified according to the side of the boundary on which it falls:

$$\boldsymbol{x} \cdot \boldsymbol{b} + n > t \Rightarrow \boldsymbol{x} \in C_1$$

The PBI model parameters are $(\boldsymbol{b}, t, \sigma)$; however the variance of the noise is assumed to be $\sigma^2 = 1$, since for any $\lambda \neq 0$, the two models given by $(\boldsymbol{b}, t, \sigma)$ and $(\lambda\boldsymbol{b}, \lambda t, \lambda\sigma)$ are identical.

3.3.4 Cue-validity models

Cue-validity models (Figure 3.8e) treat each stimulus parameter as an independent indicator of category membership, based on the relative numbers of exemplars from the two opposing categories that exhibit the *cue* (a particular value of a stimulus parameter). Thus, for example, a beard is a somewhat uncommon feature of male faces, yet it is an even less common feature of female faces, and so provides a highly valid cue to the gender category of a face.

In the weighted cue-validity model (WCVM; Reed, 1972), the validity for category C_i of the j-th parameter x_j of a test exemplar $\boldsymbol{x}$ is defined as

$$v_{ij}(\boldsymbol{x}) = p(C_i|x_j).$$

The overall cue-validity V_i is a weighted sum of these validities,

$$V_i(\boldsymbol{x}) = \sum_j \alpha_j v_{ij}(\boldsymbol{x}),$$

where the α_j are attentional weights as in the exemplar models, with $\alpha_j \geq 0$ and $\sum_j \alpha_i = 1$. Also as in the exemplar models, the decision rule incorporates Gaussian noise n and a threshold t; if the expression

$$V_1(\boldsymbol{x}) - V_2(\boldsymbol{x}) + n > t$$

is true, $\boldsymbol{x}$ is assigned to category 1, otherwise to category 2.

A modified version of this model, called the weighted frequency cue-validity model (WFCVM; Reed, 1972), uses a different definition for the validity. A weight factor,

$$q = \frac{1}{1 + F(x_m)},$$

is computed from the overall number of times $F(x_m)$ that the parameter value x_m occurs in exemplars from both categories. Then the WCVM's original validity v_{ij} is used to define the new validity

$$\tilde{v}_{ij}(\boldsymbol{x}) = \frac{1}{2} \cdot q + v_{ij}(\boldsymbol{x}) \cdot (1 - q),$$

so that the validities of rare parameter values carry little information about category membership. This reflects the idea that subjects will pay more attention to common features.

The free parameters for both the WCVM and the WFCVM are $(\boldsymbol{\alpha}, t, \sigma)$.

3.4 Model fitting

We fitted models based on several alternative representations for the schematic stimuli:

- the objects' physical parameter values (discussed in this chapter),
- the psychophysical parameter values obtained from multidimensional scaling (see Chapter 4), and
- the features derived an early vision model (Chapter 5).

Furthermore, each model could either be fitted separately to each individual subjects' data, or be fitted once to data pooled across subjects. However, since pooled fits may not accurately reflect the categorization processes of individual observers [Maddox, 1999], we used only models fitted to individual subjects' data.

Each model's free parameters were fitted to maximize its ability to predict the categorization probabilities obtained from human observers. The goodness of this fit was quantified with the *likelihood* L of the model having generated the observed probabilities, given that the fitted model correctly describes the subject's categorization process [Collett, 1991]. This likelihood is the conditional probability of the

set of observed probabilities p_i, given the values of the model parameters (which govern the predicted probabilities $\hat{p}_i$), over the N stimulus objects:

$$L = \prod_{i=1}^{N} \binom{n_i}{p_i n_i} (\hat{p}_i)^{p_i n_i} (1 - \hat{p}_i)^{(1-p_i) n_i},$$

where n_i is the number of categorization trials performed for object i, and $p_i n_i$ is the number of trials in which the observer assigned object i to category one. The likelihood takes the form of a binomial distribution because subjects' responses are treated as independent binary random variables. A numerical implementation of adaptive simulated annealing [Ingber, 1989] followed by a simplex method [Nelder and Mead, 1965] was used to maximize the likelihood L, or equivalently, minimize the minus *loglikelihood* $(-\ln L)$, which can be computed more efficiently. The range of the likelihood is $0 \leq L \leq 1$, so the range of the minus loglikelihood is $\infty \geq -\ln L \geq 0$.

We used the percentage of variance (%-variance) explained by the model as a more tangible measure for comparing fitted models. This measure is simply given by r^2, the square of the correlation coefficient between the observed and predicted probabilities.

Finally, although the loglikelihood $(\ln L)$ or %-variance are appropriate statistics for comparing fitted models having similar numbers of free parameters, comparisons of models differing in their number of free parameters, N_{fp}, require a statistic such as the Akaike information criterion [Zucchini, 2000],

$$\text{AIC} = -2 \ln L + 2 N_{\text{fp}},$$

which contains a penalty term proportional to N_{fp}. Pairwise model comparisons were made with the Wilcoxon signed-rank test of either $-\ln L$ or the AIC, and we report the median value of $-\ln L$ or the AIC to summarize the model fits from a group of individual subjects.

3.5 Model fits: Experiment 1

We found no systematic differences in the fits obtained from different model subtypes (such as those using exponential versus linear decay of similarity with distance). Therefore, in subsequent discussion, models are referred to by their general names (*e.g.*, all-exemplar models) rather than by the specific subtypes (*e.g.*, ADM

Table 3.1. Goodness-of-fit of the categorization models tested in Experiment 1.

		GCM	PBI	WPSM	WCVM
Brunswik faces {EH, ES, NL, MH}	% variance	**98.22**	98.08	96.39	88.37
	$-\ln L$	**21.15**	22.23	27.32*	42.50*
Brunswik faces {NL, MH, EH, ES}	% variance	95.68	**97.75**	95.32	74.38
	$-\ln L$	28.08	**26.53**	32.81	42.58*
Brunswik faces {MH, EH, NL, ES}	% variance	**94.02**	58.56	61.55	86.30
	$-\ln L$	**36.83**	80.57*	90.31*	52.76
cartoon faces	% variance	**95.50**	90.70	90.18	86.66
	$-\ln L$	30.68	**29.95**	37.07*	53.49
fish outlines	% variance	**97.23**	80.98	70.30	96.03
	$-\ln L$	**20.73**	32.85*	74.36*	28.74

% variance (larger value indicated better fi t)
$-\ln L$, minus loglikelihood (smaller value indicates better fi t)
bold numbers, model(s) which gave the best fi t in each row
$*$, models whose $-\ln L$ was signifi cantly worse ($p < 0.05$) than the best-fi tting model in each row

or GCM).

Table 3.1 summarizes the fits of the all-exemplar, linear boundary, prototype, and cue-validity models, for each of the five sets of objects used in Experiment 1, along with significance values for pairwise comparisons of the models using the Wilcoxon matched pair signed rank test[2]. There were two general patterns of model fits.

The first pattern was associated with the first two Brunswik face sets ({EH, ES, NL, MH} and {NL, MH, EH, ES}, which depend primarily on attention to the eyes and nose) and the cartoon faces ({EH, ES, NL, ML}). In this pattern, the all-exemplar models obtained the best fit, but the boundary model also fit well, indistinguishable from the exemplar models. The prototype models fit significantly worse ($p < 0.05$) than the all-exemplar models, but the magnitude of this difference was small. Finally, the cue-validity models fit significantly worse than the other models.

The second pattern was seen with the third Brunswik face set ({MH, EH, NL, ES}) and the fish outlines ({TF, VF, DF, MA}). As in the first pattern, the all-exemplar models obtained the best fit. However, the rest of the pattern was qualitatively different from the first pattern. Whereas the cue-validity models gave the worst fits in the first pattern, their fits were indistinguishable from the all-exemplar

[2]Note that the RXM and SPC were not used in fitting the data from Experiment 1 because even with one stored exemplar, these models carry almost as many free parameters as the number of data points to be fitted (20). This renders any comparisons among such models virtually meaningless. This issue is avoided in Experiment 2 due to the greater number of test exemplars (80).

Table 3.2. Goodness-of-fit of the models tested in Experiment 2. See also Table 3.3 for further discussion of the models' qualitative properties.

	RXM⟨1⟩	**RXM**⟨2⟩	**RXM**⟨3⟩	**SPC**⟨1⟩	**SPC**⟨2⟩	**SPC**⟨3⟩	**GCM**	**PBI**	**WPSM**
% variance	89.36*	90.98*	91.49	89.36*	90.83*	**91.64**	86.84*	87.10*	84.90*
$-\ln L$	75.72*	72.06*	71.32*	75.72*	71.65*	**69.92**	83.41*	83.66*	88.79*
AIC	**173.44**	178.13*	188.64*	**173.44**	177.30*	185.84*	178.81*	177.32	189.57*

% variance (larger value indicated better fi t)
$-\ln L$, minus loglikelihood (smaller value indicates better fi t)
AIC, Akaike Information Criterion (smaller value indicates better fi t)
bold numbers, model(s) which gave the best fi t in each row
$*$, models whose fi ts were signifi cantly worse ($p < 0.05$) than the best-fi tting model in each row

models in the second pattern. In addition, the boundary model fit very poorly, significantly worse than the exemplar models ($p < 0.05$). Finally, the prototype models fit even more poorly, significantly worse than the exemplar and boundary models ($p < 0.05$).

3.6 Model fits: Experiment 2

We fitted subjects' categorization probabilities from Experiment 2 with versions of the roaming-exemplar model and striatal pattern classifier using 1, 2, 3, 5, 7, and 10 stored exemplars[3], as well as the all-exemplar, prototype, and linear boundary models, and assessed these fits with three measures (see Table 3.2): the loglikelihood, the %-variance explained, and the Akaike information criterion (AIC).

When the model fits were assessed with their minus loglikelihoods (Table 3.2, row 2), we observed a pattern among the previously tested models similar to the first pattern observed in Experiment 1: the all-exemplar and boundary models both obtained better (lower) scores than the prototype model. However, each of these previous models was outperformed by all versions of the roaming-exemplar model and striatal pattern classifier. In addition, for both the RXM⟨n⟩ and the SPC⟨n⟩ the goodness of fit increased with the number n of stored exemplars—an unsurprising result, given that each stored exemplar reflects additional free parameters. The %-variance values (Table 3.2, row 1) show a similar pattern, but give a more concrete assessment of how well the models match the human subjects' categorization behavior: the best-fitting model (the SPC⟨3⟩) captured nearly 92%

[3]For brevity, the models with 5, 7, and 10 stored exemplars were withheld from Table 3.2, since our analysis revealed these data to merely continue the trends seen with 1, 2, and 3 stored exemplars.

of the variance, while the worst-fitting model (the WPSM) captured roughly 85% of the variance.

In contrast, when the model fits were assessed with the AIC to account for their numbers of free parameters (Table 3.2, row 3), the RXM and SPC with one stored exemplar per category (RXM$\langle 1 \rangle$ and SPC$\langle 1 \rangle$) obtained the best (lowest) scores among all models. These comparisons were statistically significant ($p < 0.05$, Wilcoxon signed rank test) except against the PBI ($p = 0.44$). Moreover, increasing the number of stored exemplars in either the RXM$\langle n \rangle$ or SPC$\langle n \rangle$ was detrimental to the AIC goodness of fit; the SPC$\langle 10 \rangle$ (AIC $= 253.29$) and RXM$\langle 10 \rangle$ (SPC $= 271.85$) fit much worse than any of the other models.

3.7 Discussion

Experiment 1 revealed a pattern of model fits similar to that reported previously [*e.g.*, Reed, 1972, Nosofsky, 1991, Maddox and Ashby, 1993, Sigala et al., 2002]. We found that across several categorization tasks involving different types of objects, an all-exemplar model provided better fits than did a linear boundary model, prototype model, or cue-validity model (Table 3.1). In some cases the fits of the linear boundary and prototype models approached those of the all-exemplar model.

The relative strengths of all-exemplar models and boundary models have been discussed at length in the literature [McKinley and Nosofsky, 1996, Maddox and Ashby, 1998, Nosofsky, 1998]. Since each model differs from the others in more than one way, it is difficult to conclude which of these differences contribute to a model's success under particular test conditions. To address this point, we introduced a "roaming-exemplar" model (RXM) that can treat independently some of the factors that were mutually dependent in previous models. It shares a flexible memory storage architecture with the striatal pattern classifier [Ashby and Waldron, 1999, Ashby et al., 2001]. It shares a decision mechanism with all-exemplar models and prototype models, since new exemplars are classified by comparing the sums of their similarities to the stored exemplars associated with each of two categories. However, in the roaming-exemplar model as well as the striatal pattern classifier, these stored exemplars are not strictly determined by the training exemplars, but are allowed to "roam" during training within the feature space of the objects to be classified.

In Experiment 2, we analyzed individual subjects' categorizations of 12 different sets of Brunswik faces by fitting them with the roaming-exemplar model and

Table 3.3. Qualitative comparison of the key models that were tested in Experiment 2.

model type	stored exemplars	main decision boundary		iso-probability contours	goodness-of-fit
		shape	*orientation*		*rank (AIC)*
linear boundary	none	linear	arbitrary	linear	2 (177.3)
prototype	1, fixed	linear	constrained	curved	4 (189.6)
roaming-exemplar$\langle 1 \rangle$	1, "roaming"	linear	arbitrary	curved	1 (173.4)
striatal-pattern$\langle 1 \rangle$	1, "roaming"	piecewise-linear	arbitrary	piecewise-linear	1 (173.4)
all-exemplar	N, fixed	curved	constrained	curved	3 (178.7)
roaming-exemplar$\langle N \rangle$	N, "roaming"	curved	arbitrary	curved	5 (279.8)

N, number of training exemplars per category
AIC, Akaike Information Criterion (smaller value indicates better fit)

striatal pattern classifier, in addition to the models used in Experiment 1 (Table 3.2). While the relationships among the all-exemplar, prototype, and linear boundary models have been analyzed previously [Nosofsky, 1990, Ashby and Maddox, 1993, Ashby and Alfonso-Reese, 1995], the improved model fits obtained with the RXM and SPC in Experiment 2 afford new insights into the strengths and weaknesses of previous models (see Figure 3.9 and Table 3.3 for an overview).

3.7.1 All-exemplar *vs.* prototype models

There are two significant differences between these models. First, in prototype models, the stored exemplars are by construction defined as the arithmetic mean in feature space of the training exemplars, while in all-exemplar models the stored exemplars occupy other locations. Second, all-exemplar models allow more than one stored exemplar per category, while prototype models allow only one, regardless of the number of training exemplars.

This second difference is linked with the question of category abstraction: storage of a category prototype implies a more abstract representation than simple memorization of all training exemplars. This places a higher burden on the learning process, since the system must select the *correct* abstraction, but makes post-learning categorization more simple, since new exemplars have only to be compared with the category prototypes. In contrast, all-exemplar models make the opposite trade-off: since no abstraction is involved, learning is straightforward as each training exemplar is simply packed away into memory, but post-learning categorization is complicated since a new exemplar must be compared with every stored exemplar in memory. While this requirement is not neurobiologically unreasonable in typical psychophysical experiments which use few training ex-

emplars per category, it seems less likely to be applicable to natural visual categories, which may contain thousands or more of exemplars. Furthermore, biological systems are likely to spend more time in using categories than in learning them, at least for highly salient categories (*e.g.*, male/female faces, poisonous/non-poisonous fruit). Such arguments lend some *a priori* credence to the notion of a prototype model, but are entirely hidden from statistical comparisons, since neither the *contents* of the memory nor the complexity of the learning process are free parameters of the models. Indeed, past comparisons between all-exemplar and prototype models have generated a preponderance of evidence favoring the all-exemplar models.

When the contents of the memory locations become explicit free parameters, questions concerning the importance of memory capacity can be addressed statistically. For example, by comparing either the RXM$\langle 1 \rangle$ or the equivalent SPC$\langle 1 \rangle$ with a prototype model, we examine only the first difference mentioned above between all-exemplar models and prototype models (whether memory traces are fixed at the category mean). On the other hand, by comparing the RXM$\langle 1 \rangle$ with the RXM$\langle n \rangle$ ($n > 1$) we examine only the second difference (changing the number of stored exemplars). Our results from Experiment 2 (Table 3.2) demonstrate a large improvement from allowing roaming, rather than fixed, stored exemplars (AIC: RXM$\langle 1 \rangle$, SPC$\langle 1 \rangle$ = 173.4, prototype = 189.6), while allowing additional stored exemplars actually leads to a decline in goodness-of-fit when the additional memory is counted among the models' free parameters (AIC: RXM$\langle 10 \rangle$ = 271.9, RXM$\langle 1 \rangle$ = 173.4). Thus, although the empirical success of all-exemplar models appears to support a rejection of category abstraction, our results show that in fact we should only reject the strict notion of abstraction involving category prototypes.

3.7.2 Prototype *vs.* linear boundary models

These two models are similar in that each has a *decision boundary* (i.e., the iso-probability density surface where the categorization probability density equals 0.5) that is a hyperplane in stimulus parameter space [Ashby and Maddox, 1993]. The models also have two important differences. First, for prototype models, the decision boundary must be orthogonal to the vector connecting the two category prototypes in stimulus parameter space, while for linear boundary models, the decision boundary can have an arbitrary orientation. Second, consider the iso-probability density surfaces with $p \neq 0.5$: for the linear boundary model, these

are hyperplanes parallel to the decision boundary, but for the prototype model, these are paraboloid surfaces with a curvature that increases as p diverges from 0.5 (see Figure 3.9). Conceptually, this means that for the linear boundary model, decision thresholds are the same at every point along the category boundary in feature space, while for the prototype model, decision thresholds are narrowest (*i.e.*, the model is most confident) at the center of feature space, near the category prototypes. Intuitively, the behavior of the prototype model seems more natural—new objects are categorized more accurately when they are similar to previously seen objects—but our results from Experiment 1 along with others' results [*e.g.*, Nosofsky, 1991] clearly contradict this intuition.

Again, a more flexible model can help to provide some insight into this issue. In particular, the RXM$\langle 1 \rangle$ and SPC$\langle 1 \rangle$ are like the prototype model with curved, rather than planar, iso-probability surfaces, but are like the linear boundary model in that the main decision boundary can have an arbitrary orientation. Our results from Experiment 2 demonstrate that with these two qualities combined, the RXM$\langle 1 \rangle$ and SPC$\langle 1 \rangle$ fit human behavior significantly better than either the prototype or linear boundary models (AIC: RXM$\langle 1 \rangle$, SPC$\langle 1 \rangle$ = 173.4, prototype = 189.6, linear boundary = 177.3).

3.7.3 All-exemplar *vs.* linear boundary models

By extension of the previous two comparisons, the differences between the all-exemplar model and the linear boundary model are even more numerous. The all-exemplar model allows for curved decision surfaces, but the orientation of the surface has limited flexibility. In contrast, the linear boundary model allows only flat decision surfaces, but these may have arbitrary orientation. Again, the RXM can combine the separate strengths of these two models.

In the RXM, the parameters which describe the stored exemplars become free parameters of the model, and can be incorporated into comparisons among models using statistical measures such as the Akaike Information Criterion. This allows us to address the importance of memory by comparing different versions of the RXM with different numbers of stored exemplars. With this framework, we can now provide a better answer as to why models which are otherwise appealing in their conceptual simplicity, such as prototype models, are consistently outperformed by all-exemplar models: all-exemplar models allow better flexibility in matching the shape and orientation of decision surfaces to those used by human

observers. Our results show that the goodness-of-fit of all-exemplar models can be improved upon by allowing "roaming" stored exemplars, and thus an unconstrained decision boundary, without committing to high memory demands or to a lack of category-level abstraction.

3.7.4 RXM *vs.* SPC

Computationally, the RXM and SPC are quite similar to each other, as well as to several earlier models (Anderson, 1991, Kruschke, 1992; see also Ashby and Waldron 1999), in that they each rely on a set of units representing locations in feature space, and categorize new inputs based on the distance in feature space between the input and the various stored units. The main qualitative difference is at the decision stage, where the RXM produces smoothly curved decision boundaries, while the SPC produces piecewise-linear decision boundaries. This is because in the RXM, the categorization decision is based on contributions from all of the stored units, with weights proportional to the distance of the stored units from the input, while in the SPC, only the nearest stored unit of each category is considered. In this sense, the SPC involves a much stronger nonlinearity than the RXM. This sharp nonlinearity may not be strictly implemented in neural circuitry; rather, a biological implementation might have to rely on a "softmax" approximation [Riesenhuber and Poggio, 1999] which would more closely resemble a gradual decay of similarity with distance as in the RXM. This question remains to be resolved by further neurophysiological study.

Despite the computational similarities of the SPC and RXM, each is derived from previous models whose neurobiological implications may appear to put the two models at odds. We have proposed the RXM as a generalization of prototype models and all-exemplar models (i.e., GCM). All-exemplar models, which propose one hidden unit for every training exemplar, have in particular carried the implication that observers may rely on explicit memory of individual visual stimuli, and that the models' hidden units correspond to these memory traces [*e.g.*, Knowlton, 1999]. This stands in contrast to neuropsychological evidence from patients with amnesia who, despite an impairment in recognition tasks requiring declarative memory of individual exemplars, are relatively unimpaired in various tasks requiring category learning [Knowlton and Squire, 1993, Squire and Knowlton, 1995, Filoteo et al., 2001]. For this reason, Ashby and Waldron [1999] proposed that the striatal units in the SPC are primarily response-associated; that is, the

units are primarily involved in decision, rather than perception. We do not make any claims regarding whether the hidden units in the RXM are essentially explicit memory traces, particularly since the hidden units are allowed to occupy points in feature space that were never directly related to a training exemplar. However, electrophysiogical evidence does suggest that the mechanisms that are shaped during category learning also affect perception. Sigala and Logothetis [2002] showed that, after category learning, inferotemporal neurons in the macaque were more sensitive to features that were diagnostic of category membership than to non-diagnostic features (although Ashby and Ell [2001] reviewed studies in which exposure to visual stimuli that were associated with *non-visual* categories such as good/bad tastes did *not* lead to a change in visual cell response properties). Furthermore, behavioral data (MDS) showed that monkeys' *perception* also shifts as a result of category training [Sigala et al., 2002], supporting the idea that the hidden units tuned to specific features in a categorization model may not operate solely at the decision stage, but may also be directly involved in perception. This is not incompatible with the evidence from amnesic patients; it may be that categorization relies on neural representations that are explicit in the sense of being discrete and minimally distributed, but do not constitute "explicit memory" in the sense of being behaviorally accessible for declarative memory. In any case, current psychophysical evidence alone cannot discriminate whether a model's mathematical constructs correspond to neuronal processes occurring in specific cortical areas such as the striatum, inferotemporal cortex, or even prefrontal cortex.

3.7.5 Generalization and learning

In the most general terms, categorization is a process with four components: (1) external input (visual stimuli), (2) internal input (pre-existing memories and neural state), and (3) a mechanism that combines the inputs to produce (4) an observable output (categorization behavior). A complete theory of categorization should quantitatively describe an internal mechanism that can be appropriately tuned by a learning process involving exposure to a limited set of training exemplars [*e.g.*, Nosofsky et al., 1992, Ashby and Ell, 2001], and should describe how differences in observers' pre-existing internal states lead to different categorization behavior given the same input. In the context of the RXM or SPC, for example, such a theory might help address questions such as how the number of hidden units is adjusted during learning, perhaps in relation to the difficulty in separating categories from

one another.

By this standard, the models we have discussed provide only a partial theory, in that they only describe the fully-trained mechanism without offering a process for learning the tunable parameters of that mechanism. We have inferred the final values of these parameters by fitting the models to human behavior on a set of test exemplars[4]. In other words, by collecting and modeling observers' responses to the test exemplars, we have only addressed the question of *what did observers learn*, rather than the more complex question of *how did they learn it*. Nevertheless, our descriptive results provide valuable constraints for more complete future models of the learning process; after all, a model cannot successfully describe the learning process without also successfully describing the outcome of that process.

[4]An alternate approach would be to fit the models to match observers' performance on the training set, and then judge the models based on observers' performance on the test set. But our observers were trained to be highly accurate in categorizing the training set (with most categorization probabilities near 0 or 1), so their training-set performance places only very weak constraints on the models, since all of the models can be trivially fitted to classify the training set with 100% accuracy. In contrast, we designed the test exemplars for the express purpose of being potentially ambiguous, so that observers' test-set performance would place strong constraints on the models being fitted.

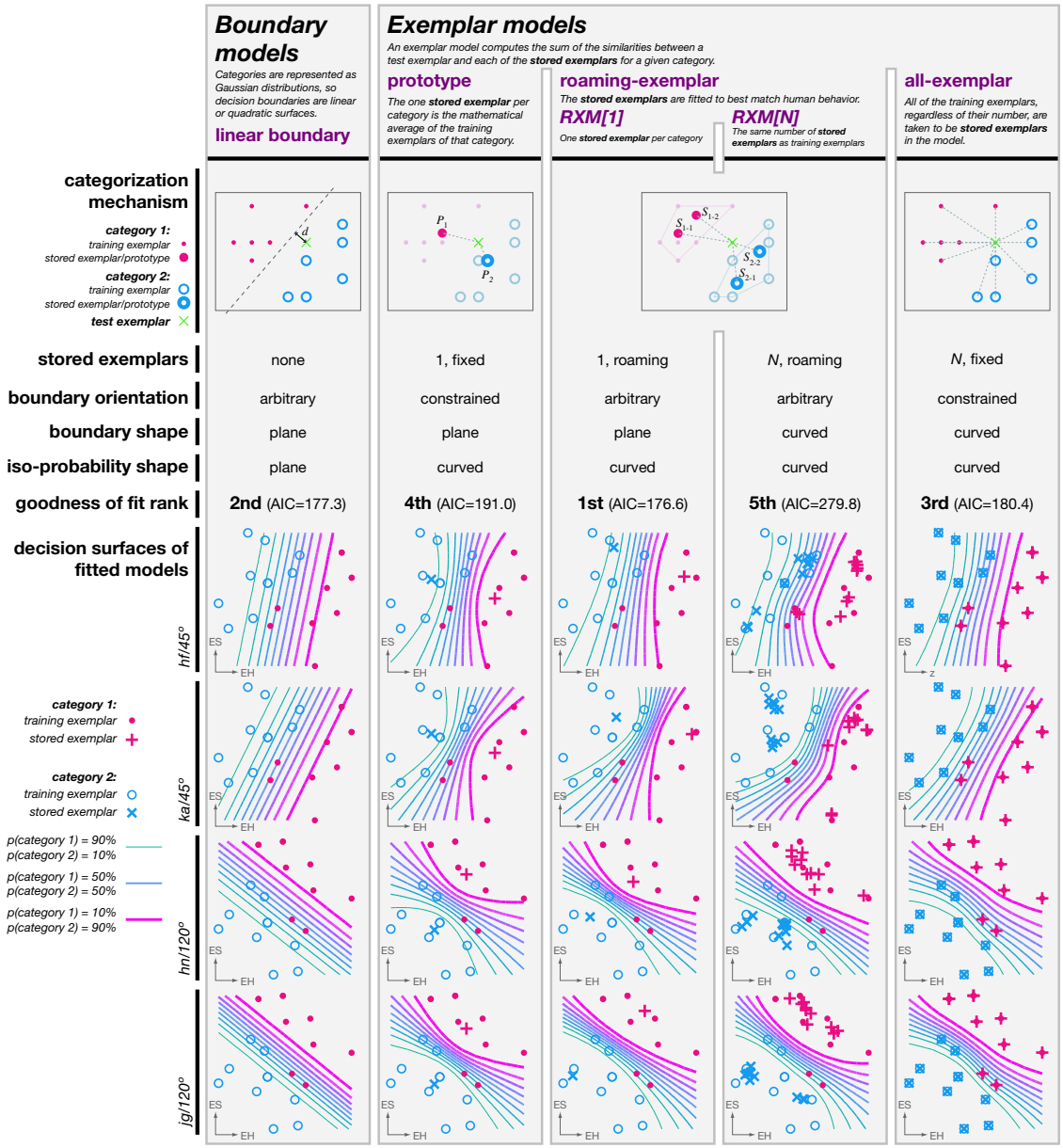


Figure 3.9. An illustration of the type of decision surfaces predicted by the various models. Each column summarizes the behavior of one of the model types. The top row reproduces the schematic diagrams of the models from Figure 3.8. The next rows summarize some of the qualitative features of each model. Finally, the bottom rows show the decision surfaces that resulted when each model was fitted to one subjects' categorization behavior (one subject per row). Each of the contour lines through these plots represents an iso-probability contour, along which the model predicts a constant categorization probability. Nine consecutive iso-probability contours are shown for category 1 probabilities of 10–90%. The central iso-probability contour represents the main decision boundary, or equivocation point, where the model predicts that the subject would be equally likely to categorize a stimulus into either category 1 or category 2. Note how the different models produce iso-probability contours with different characteristic shapes.

CHAPTER 4

Multidimensional Scaling

4.1 Introduction

Current categorization models in the psychological literature, such as those presented in the previous chapter, typically depend on high-level multidimensional representations of incoming stimuli [Ashby, 1992b, Ashby and Maddox, 1993]. Edelman [1999] reviewed evidence suggesting that such representations are intimately linked with the perceptual similarity of stimuli. A common technique used to infer implicit psychological representations is to apply *multidimensional scaling* (MDS) to observers' judgments of similarity about a set of stimuli. Presently, the link between these psychophysical measures of similarity and the neuronal mechanisms underlying stimulus representation in the primate visual system remains poorly understood. A number of new approaches, including functional brain imaging in humans [Edelman, Grill-Spector, Kushnir, and Malach, 1998] and electrophysiological recordings in trained macaque monkeys [Op de Beeck et al., 2001, Sigala and Logothetis, 2002] are likely to shed light on these issues. Some of the methods are only applicable to one species, yet we would ultimately like to draw conclusions that generalize (at least to higher primates); therefore a final understanding must rely on comparisons between inferred psychological representations in monkey and human observers. Since it is nearly impossible to train animals to give graded similarity ratings between pairs of objects (the common method in human studies), animal studies must rely on simpler two-alternative forced choice methods instead. Our aim therefore is to directly compare these two

ways of rating object similarity directly in human subjects, with the goal of establishing a link that can support future cross-species comparisons.

4.2 Similarity tasks

Two different similarity tasks (pairs and triads tasks, see Figure 4.1) were performed with the 20-object configurations used in Experiment 1 from Section 3.2 (Figure 3.1). For each configuration, subjects' psychophysical responses were used to form a 20×20 experimental *dissimilarity matrix* with entries δ_{ij}, using a procedure specific to the task (see descriptions below). This matrix was then used to estimate subjects' psychological representations of the stimuli (see Section 4.3).

4.2.1 Pairs task

In the pairwise comparison task [Borg and Groenen, 1997, ch. 6.2], or *pairs task* (Figure 4.1a), subjects viewed sequences of simultaneously presented pairs of objects. Each pair was presented for 2 s, followed by 2 s of blank screen. Subjects could respond at any time during that 4 s interval with a button press between 1–9, indicating how similar the objects appeared. Subjects were instructed to choose "9" if and only if the two objects were identical. Each of the 400 possible pairings of the 20 objects was presented 3 times throughout the experiment, giving 1200 total trials. For each pair of objects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, the dissimilarity matrix entry δ_{ij} was taken to be $9 - \bar{s}_{ij}$, where $\bar{s}_{ij}$ is the average similarity rating over the n trials containing objects i and j, with $n = 3$ for $i = j$, $n = 6$ for $i \neq j$ because pairings with distinct objects were shown three times for each of the two possible orderings $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $(\boldsymbol{x}_j, \boldsymbol{x}_i)$.

As a measure of the consistency with which subjects rated the pairs, we used a triangle inequality test. For each triplet of objects $(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k)$, we considered the three corresponding pairwise ratings given by a subject, $\bar{s}_{ij}$, $\bar{s}_{ik}$, and $\bar{s}_{jk}$. A strict triangle inequality rule would require

$$\begin{aligned} \bar{s}_{ij} + \bar{s}_{ik} &> \bar{s}_{jk} \\ \bar{s}_{ij} + \bar{s}_{jk} &> \bar{s}_{ik} \\ \bar{s}_{ik} + \bar{s}_{jk} &> \bar{s}_{ij} \end{aligned}$$

However, to allow for the possibility that subjects might place different objects at

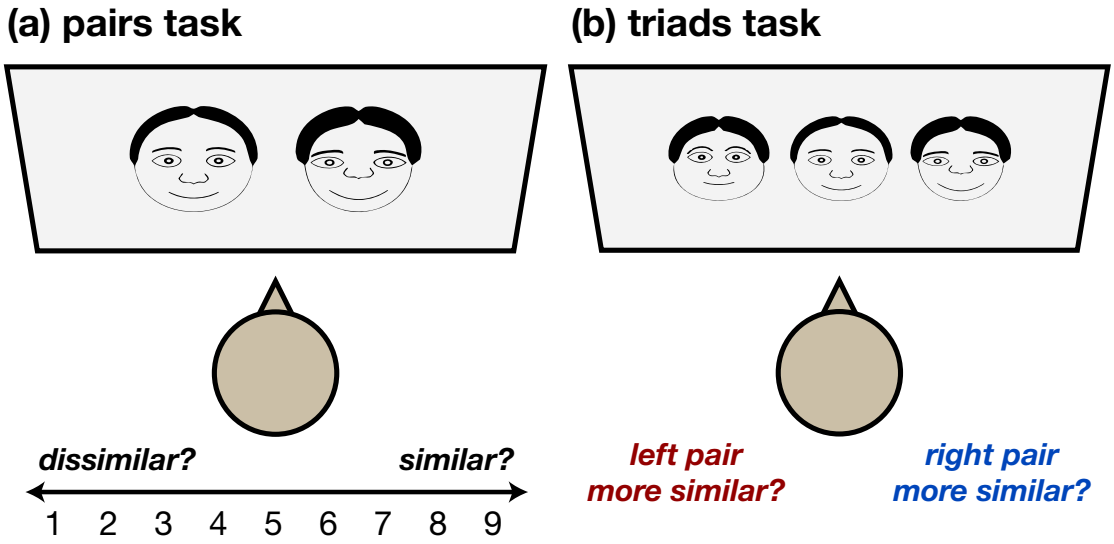


Figure 4.1. The two psychophysical tasks shown here were used to acquire similarity data that could drive a *multidimensional scaling* (MDS) analysis. In the *pairs task* **(a)**, subjects viewed successive pairs of objects on a computer screen. After each pair, the subject was required to respond with a rating from 1 to 9 indicating how similar the two objects seemed to each other, with 9 meaning "identical." In the *triads task* **(b)**, subjects viewed successive triads of objects; for each triad, subjects were required to indicate whether the left or right pair seemed more similar. The triads task, but not the pairs task, can be learned by monkeys, while human subjects prefer the pairs task since it requires fewer trials. With either task, the raw data are transformed into a dissimilarity matrix whose entries indicate the perceived dissimilarity between each pair of objects; this matrix is then used to guide the construction of a low-dimensional representational space for the objects such that the inter-object Euclidean distances in this space correlate well with the observed dissimilarities.

identical locations in perceptual space, we tested a relaxed triangle inequality rule:

$$\begin{aligned} \bar{s}_{ij} + \bar{s}_{ik} &\geq \bar{s}_{jk} \\ \bar{s}_{ij} + \bar{s}_{jk} &\geq \bar{s}_{ik} \\ \bar{s}_{ik} + \bar{s}_{jk} &\geq \bar{s}_{ij} \end{aligned}$$

Across 20 pairs task experiments, with 1540 possible triangles per experiment, we found that 96.05% (29583 of 30800) satisfied this relaxed triangle inequality. On the other hand, 91.43% (28160 of 30800) satisfied the strict inequality.

4.2.2 Triads task

The *triads task* (Figure 4.1b), a variant of the anchor stimulus method [Borg and Groenen, 1997, ch. 6.2], is a two-alternative forced-choice (2-AFC) task, and as

such it has been particularly useful for studies involving non-verbal observers (*e.g.*, human infants, Arabie, Kosslyn, and Nelson, 1975; monkeys, Sigala et al., 2002). Subjects viewed sequences of simultaneously presented triads of objects, arrayed horizontally. Each triad $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$ was presented for 2 s, followed by 2 s of blank screen. Subjects could respond at any time during that 4 s trial with a button press indicating whether the left pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ or the right pair $(\boldsymbol{x}_2, \boldsymbol{x}_3)$ appeared more similar. Time constraints prohibited using all possible triad combinations. Instead, the 6840 possible triads $(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k)$ of the 20 objects were sorted by the Euclidean distance in stimulus parameter space between the leftmost and rightmost stimuli ($d(\boldsymbol{x}_i, \boldsymbol{x}_k)$), and the 1710 triads with the largest such distances were used for psychophysics. Finally, subjects' binary responses in the triads task were transformed into analog dissimilarities δ_{ij} using a procedure described in Sigala et al. [2002].

4.3 MDS analysis

Multidimensional scaling (MDS) was used to find a set

$$\hat{\boldsymbol{X}} = \{\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_N\}$$

of N 4-D vectors $\hat{\boldsymbol{x}}_i$, that best reflected the internal psychological representation used by a subject when performing a similarity task[1]. The best such representation is found by minimizing the *stress* σ,

$$\sigma = \frac{1}{2}\sum_{i,j}(d(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_j) - \delta_{ij})^2,$$

where d is the Euclidean distance and δ_{ij} are the dissimilarities computed from subjects' responses in one of the similarity tasks. These representations allow for a clear correspondence between the scaled dimensions and the physical stimulus parameters, as follows.

To align the MDS configuration $\hat{\boldsymbol{X}}$ with the original configuration $\boldsymbol{X}$, we used an isometric *Procrustes transformation* P, consisting of a rigid rotation, translation, and uniform scaling [Borg and Groenen, 1997, ch. 19]. The optimal Procrustes

[1]Note that this procedure deviates from a strict definition of MDS because the dimensionality of the representation space was fixed to 4, rather than being a free parameter. However, previous studies using have obtained satisfactory MDS solutions with 4-D representations Brunswik faces [Nosofsky, 1991] and fish stimuli [Sigala et al., 2002].

transformation $P_{\min}$ minimizes the loss function

$$L(P) = \sum_i d^2(\boldsymbol{x}_i, P(\hat{\boldsymbol{x}}_i)).$$

This minimum value $L(P_{\min})$, called the *residual squared distance* (RSD), quantifies the dissimilarity between subjects' psychological representation $\hat{\boldsymbol{X}}$ and the original stimulus configuration $\boldsymbol{X}$.

To determine whether the observed RSDs were smaller than would be expected by chance, a Monte Carlo technique was used. RSDs were computed between the original configuration and 10^5 random configurations whose parameters were drawn from a uniform distribution over $[0, 1]$. The resulting distribution was used to estimate the significance levels of the RSDs of the pairs and triads MDS configurations.

4.4 MDS results

In order to quantify the goodness of fit between subjects' Procrustes-transformed MDS configurations and the original stimulus configuration, we used Monte Carlo simulations comparing the residual squared distances (RSDs) of our subjects' MDS configurations with the RSDs of random configurations (see Figure 4.2a). Figure 4.3 shows the alignment between one example original stimulus configuration and the corresponding Procrustes-transforms for each of

- a random configuration from the Monte Carlo simulation,
- an MDS configuration derived from the triads task, and
- an MDS configuration derived from the pairs task.

The mean of the pairs-MDS distribution (0.1444) was roughly twice as close to the original configuration as would be expected by chance (0.3268), and all pairs-MDS configurations were significantly closer ($p < 0.005$) to the original configuration than were the random configurations. Likewise, the triads-MDS configurations were also all significantly closer to the original space than would be expected by chance ($p < 0.005$), although the mean of the triads-MDS distribution (0.1982) was not as close to the original configuration as was the pairs-MDS distribution. A paired t-test showed that the residual squared distances of the pairs-MDS config-

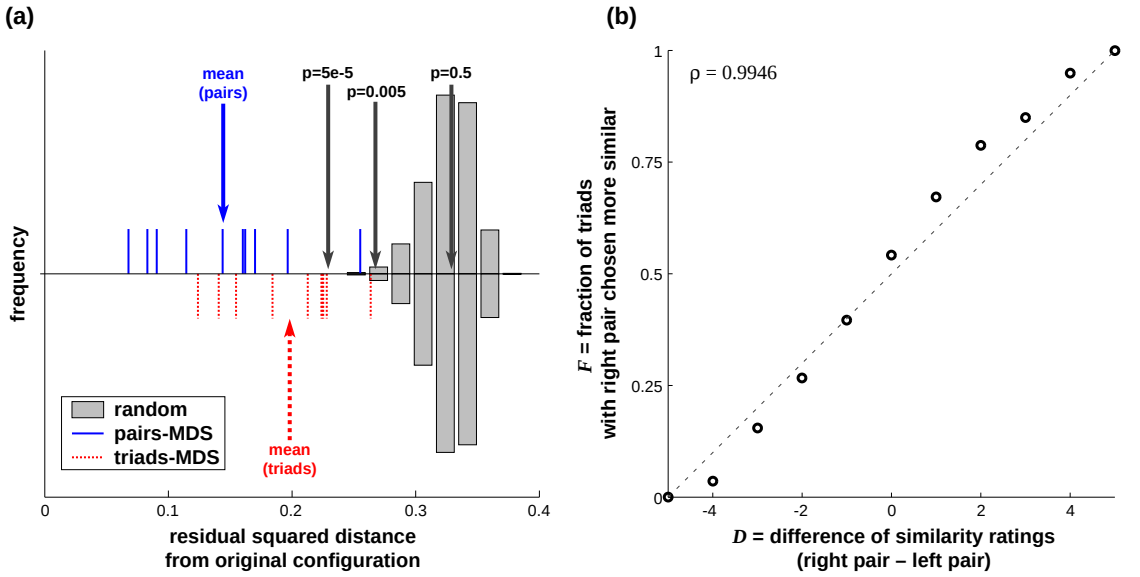


Figure 4.2. A summary the MDS configurations obtained with pairs and triads similarity tasks. **(a)** As measured with the *residual squared distance* (RSD), all of the pairs-MDS and triads-MDS configurations were significantly more similar ($p < 0.005$) to the original configuration of stimulus parameter values than would be expected by chance. The distribution of RSDs for 10^5 random configurations (gray bars, arrows with p-values) was compared with the RSDs for 10 subjects' pairs-MDS (upper, solid lines) and triads-MDS (lower, dashed lines) configurations. Two identical configurations would give an RSD of 0, while two unrelated configurations would give an RSD near the median of the random distribution (0.33). The RSDs for pairs-MDS were significantly smaller than those for triads-MDS ($p < 0.05$). **(b)** To directly compare the similarity judgments obtained in the pairs and triads tasks, we computed two metrics for triads of objects $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$: (1) the difference $D = S(\boldsymbol{x}_2, \boldsymbol{x}_3) - S(\boldsymbol{x}_1, \boldsymbol{x}_2)$ of two similarity ratings given *in the pairs task*, and (2) among triads with similar values of D, the fraction F of trials in which the observer chose $(\boldsymbol{x}_2, \boldsymbol{x}_3)$ as more similar than $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ when viewing $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$ *in the triads task*. The two measures D and F were highly correlated ($\rho = 0.9946$) across 10 subjects.

urations were significantly smaller than those of the triads-MDS configurations ($p < 0.05$).

To further assess the relationship between these two methods for obtaining similarity judgments, we performed a more direct comparison, using subjects' raw responses rather than the derived MDS configurations (Figure 4.2b). In each trial in the triads task, subjects viewed three objects $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$ and compared the similarities of the two pairs $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and $(\boldsymbol{x}_2, \boldsymbol{x}_3)$. Subjects also directly rated the similarities of these pairs in the pairs task. Thus, for each triad $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$ which was shown in the triads task, we computed D_{pairs}, the difference between the similarity ratings given by the subjects *in the pairs task* to the pairs $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and $(\boldsymbol{x}_2, \boldsymbol{x}_3)$. We then split the triads trials into groups with similar values of D_{pairs}. Within each group

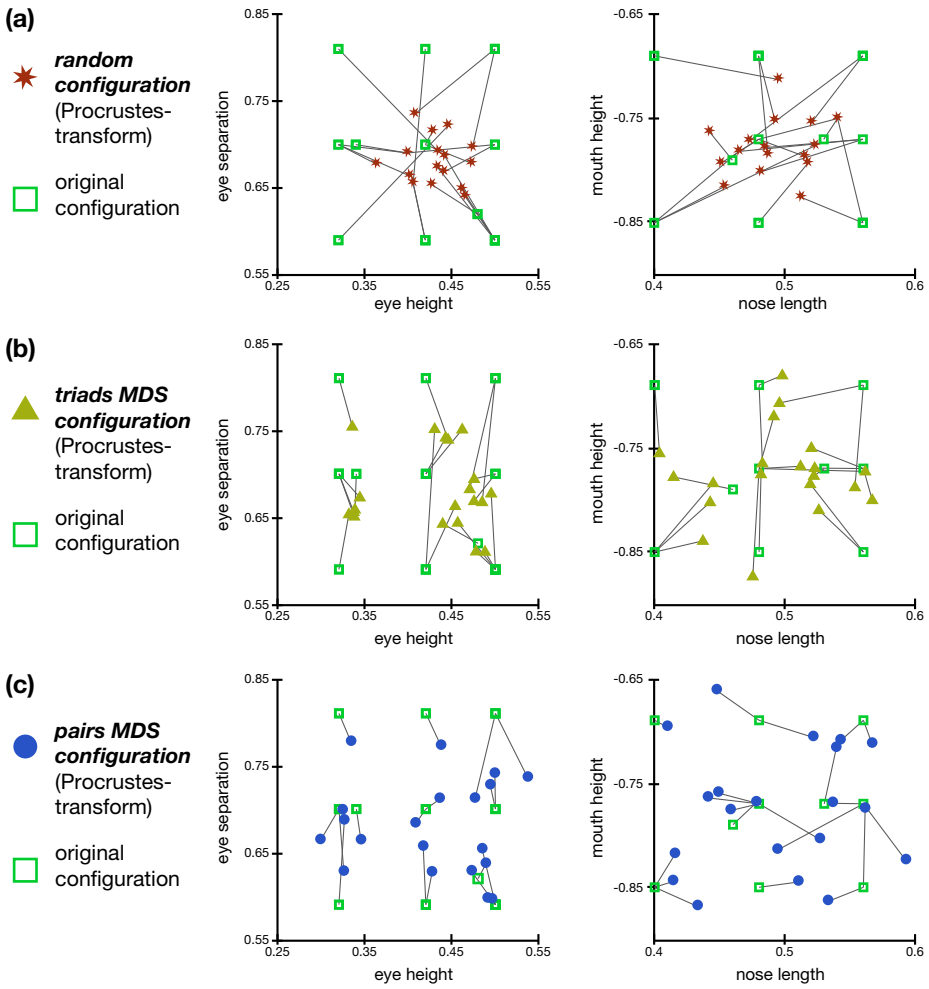


Figure 4.3. Shown here is the alignment between the original stimulus configuration and MDS configurations for one sample set of Brunswik faces. The location of each Brunswik face within the original stimulus configuration is given by a pair of points, one in each of a pair of plots, with two of its feature values indicated in the left plot (eye height and eye separation), and the other two feature values indicated in the right plot (nose length and mouth height). Each pair of plots shows the relationship between the original stimulus values (□) and the feature values derived from MDS; a thin line is shown connecting these alternate representations for each Brunswik face. **(a)** The alternate representation here (∗) is actually a configuration in which each feature value was assigned randomly, and the resulting configuration was then aligned as best as possible to the original configuration with a Procrustes transformation. **(b)** Here, the alternate representation (▲) is derived via MDS from one subject's responses in the triads task, and the resulting configuration is again Procrustes-transformed to match the original configuration as best as possible. **(c)** Finally, here the alternate representation (●) is derived via MDS from the same subject's responses in the pairs task. In the examples shown here, the apparent trend is that the pairs-MDS configuration is more similar to the original than is the triads-MDS configuration, which in turn is more similar to the original than is the random configuration. In fact, this trend carries over to the overall results shown in Figure 4.2.

we computed F_{triads}, the fraction of trials for which the subject chose the right pair as more similar than the left pair *in the triads task*. These two measures D_{pairs} and F_{triads} were highly correlated in data obtained from single subjects ($\rho > 0.98$ for 9 of 10 subjects) and when data were pooled across subjects ($\rho = 0.9946$; Figure 4.2b).

Finally, for each of the models tested in Experiment 1 (Chapter 3), we re-fitted the model using the pairs- and triads-derived MDS configurations in place of the original stimulus configurations. Thus another indicator of the relationship between the psychophysical (*i.e.*, MDS) and physical features is given by the relationship between the goodness of fit of models fitted with those two sets of features. Measured by the % of variance explained, both the MDS-pairs and MDS-triads model fits were strongly correlated with the fits obtained using the original configuration, as well as with each other ($\rho > 0.90$ in each case). The average goodness of fit of the MDS-pairs models lagged behind that of the original models by 2.3 %-variance, and the MDS-triads models lagged by an additional 5.5 %-variance. Figure 4.4 shows these relationships for the MDS-pairs configurations.

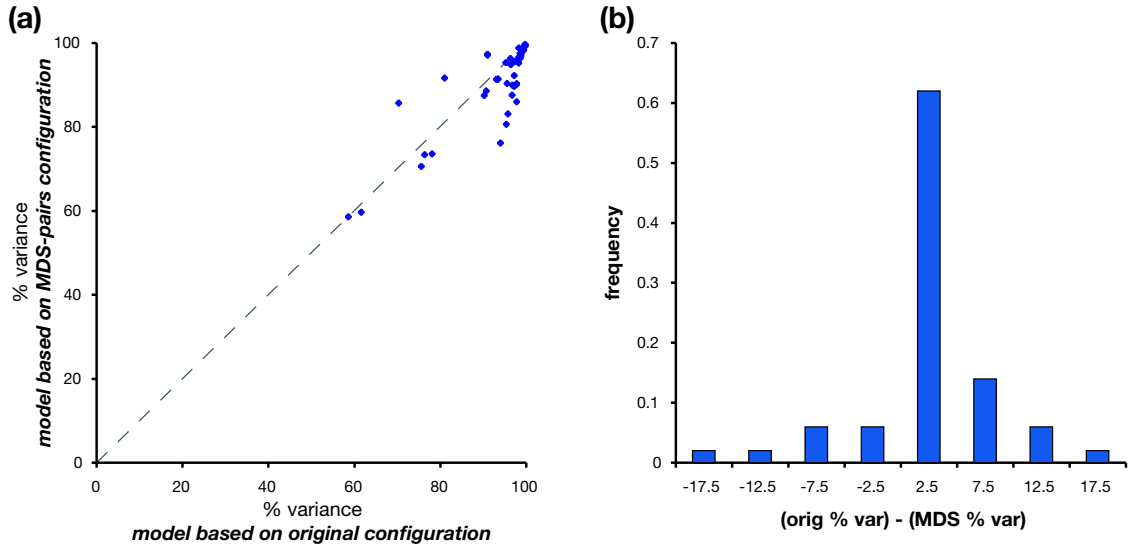


Figure 4.4. **(a)** A scatter plot showing the relationship between the goodness of fit (% of variance explained) of categorization models based on the original stimulus configurations (*i.e.*, the physical feature values like eye height and eye separation) and the fit of models based on the MDS configurations derived from subjects' responses in the pairs task. The dashed diagonal line indicates where the two fits would be equal; it can be seen that the majority of the points fall just below this line, indicating that the models based on the original configurations fit slightly better than those based on the MDS configurations. This trend is seen more explicitly in **(b)**, which shows a histogram of the paired differences in the goodness of fit between the two types of model. In the majority of cases, the model based on the original configuration has a 2.5% advantage in %-variance over the MDS-based model.

4.5 Discussion

Several authors [Shepard, 1987, Edelman, 1999] have proposed that neural mechanisms of representation are based on similarity. Similarity measures can be transformed to feature space representations with multidimensional scaling, a technique that has often been used as the basis for models of categorization and recognition [*e.g.*, Nosofsky, 1986]. Yet, only recently has the neurobiological validity of MDS begun to be investigated directly with monkey electrophysiology [Op de Beeck et al., 2001, Sigala and Logothetis, 2002] and human fMRI [Edelman et al., 1998]. Given the practical significance of comparing results obtained in monkey and human studies, it is important to establish the compatibility of the behavioral methods used for the two species. Because it is impossible for monkey observers (as well as for human infants; *e.g.*, Arabie et al., 1975, Sloutsky and Lo, 1999) to give an analog similarity rating, a task based on binary choice such as the "triads" task must be used instead[2]. Unfortunately, since each triads trial conveys only relative information about pairwise similarities, the entire task requires many trials and is quite time demanding. Thus, adult human subjects prefer the "pairs" task, which is is based on analog similarity judgments, and is less time demanding since each trial directly conveys absolute information about pairwise similarities. Therefore, we compared the results of the pairs and triads tasks within a set of human subjects to assess their equivalence in characterizing psychophysical representations of similarity. As Figure 4.2b shows, the judgments obtained in theses two tasks were highly correlated, suggesting that a shared process could account for subjects' performance in both tasks. These results legitimize comparisons between data from the pairs task in human subjects and data from the triads task in monkey subjects.

One purpose of the MDS analysis is to construct an input representation for the categorization models that can be tested independently of the original stimulus configuration. We found that model fits did not improve when the models were based on pairs-MDS or triads-MDS configurations, relative to the original stimulus configuration (Figure 4.4). This result agrees with the findings of Sigala et al. [2002] using both monkey and human subjects in experiments similar to those reported here. Thus, although some models (such as the GCM; Nosofsky, 1986, 1991) have originally been used exclusively with MDS configurations, we found that

[2] Alternatively, a same/different task can be used to generate a confusion matrix for MDS [Sugihara, Edelman, and Tanaka, 1998].

they achieve similar performance when the original configuration is used instead. We interpret these results to mean that subjects can efficiently learn a psychological representation that is highly similar to the native representation of a set of objects. The mechanism for this learning process remains a subject for future investigation, and the next chapter explores one model of early vision for extracting such a representation directly, starting with the visual input at the retina. In any case, the empirical correlation between the original and MDS configurations is of practical relevance because the MDS procedure is time-intensive both in the collection of similarity task data and in the computational analysis of those data. Our results suggest that this analysis step can be bypassed without affecting the comparison of various classification models.

CHAPTER 5

Early vision in categorization

In a biological system, any high-level representation must be built from lower-level representations, and in vision this means that all representations must ultimately trace back to the retinal input. Many categorization models presuppose that the high-level (external) features used by the experimenter to define the objects are the same as those used internally by the observer when making a categorization decision. For example, many categorization studies have used a set of circles with bisecting lines, defined by two features: the diameter of the circle, and the angle of the bisecting line (see Figure 2.1). This approach has certainly been fruitful, and MDS studies (Chapter 4) have demonstrated strong similarities between the external and internal feature representations. Nevertheless, apparent irregularities in the categorization process that might be inexplicable in terms of high-level representations, could appear entirely natural in the light of biological early vision. At the least, features such as *angle of the bisecting line* are not likely to be represented explicitly by neurons involved in visual perception; rather, a population of neurons might form a distributed representation, in which each neuron responds preferentially to a single range of orientations. Whether such differences have an effect on the output of categorization models is an empirical question. We have tested a set of hybrid models, in which we adapted a hierarchical model of early vision model ("HMAX") based on Riesenhuber and Poggio [1999]. HMAX operates directly in image space, in contrast to the categorization models described above, which operate in feature space. Our approach was to extract a new feature space representation from the output of HMAX, which could then be used as an

alternate input for fitting the categorization models, to be compared with model fits obtained using the original physical feature space.

5.1 HMAX: a model of early vision

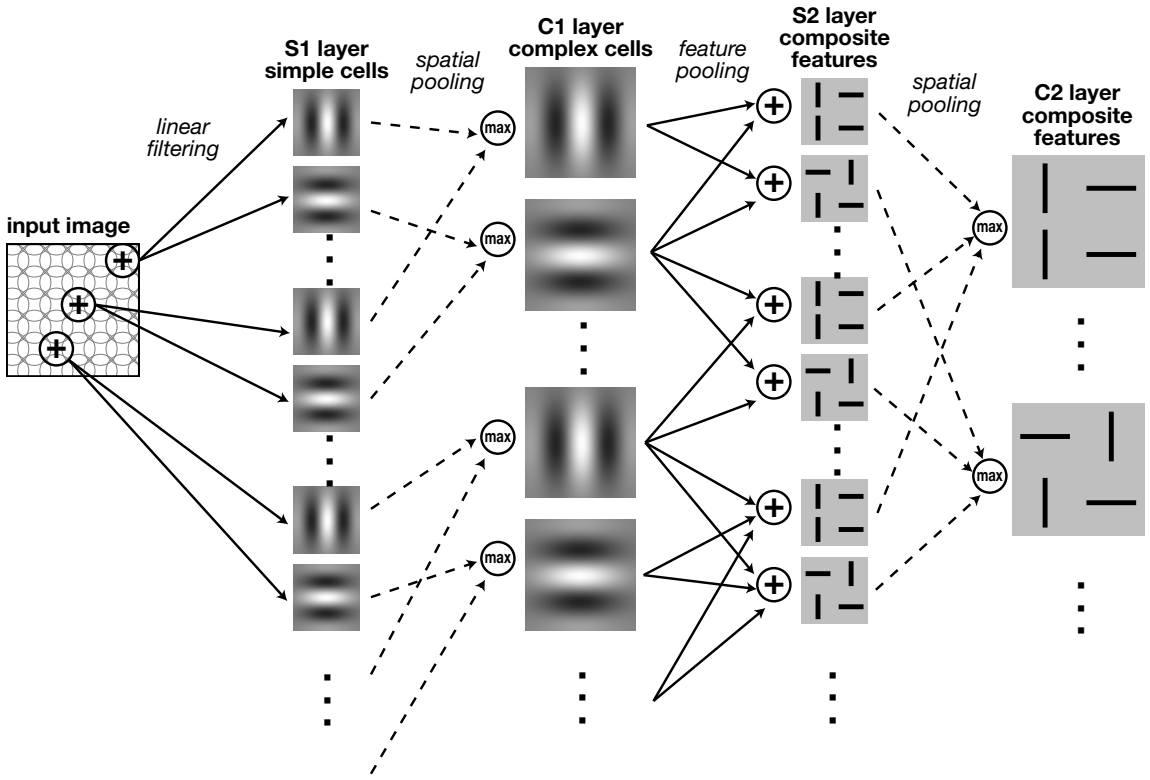


Figure 5.1. Shown here is a schematic depiction of the "HMAX" model of early vision [Riesenhuber and Poggio, 1999]. The workings of the model progress from left to right in this diagram. First, an input image is filtered with overlapping orientation-tuned linear filters at multiple orientations and spatial scales. After additive pooling over small regions of spatial locations, these filter outputs form the S1 layer responses (next column). Next, the outputs from the S1 layer are pooled over spatial locations and spatial scale bands, this time using a *max* operation, to form the C1 layer. In the next step, more complex features are formed, in which units in the S2 layer respond to different spatial arrangements of oriented edges. Finally, the C2 layer is formed by pooling S2 responses over spatial locations and spatial scale bands. In the original model, C2 cells pooled over *all* spatial locations and *all* scale bands; however, in our modified model (see Figure 5.5) the C2 cells pooled only over coarse sub-regions of the entire image.

In brief (see Figure 5.1), HMAX operates through two stages of "simple" and "complex" units (S1, C1, S2, and C2). The S1 representation (see Figure 5.2 for an example) is obtained by filtering the image with a bank of Gabor-like filters tuned for multiple orientations and spatial scales. The C1 representation (Figure 5.2) is

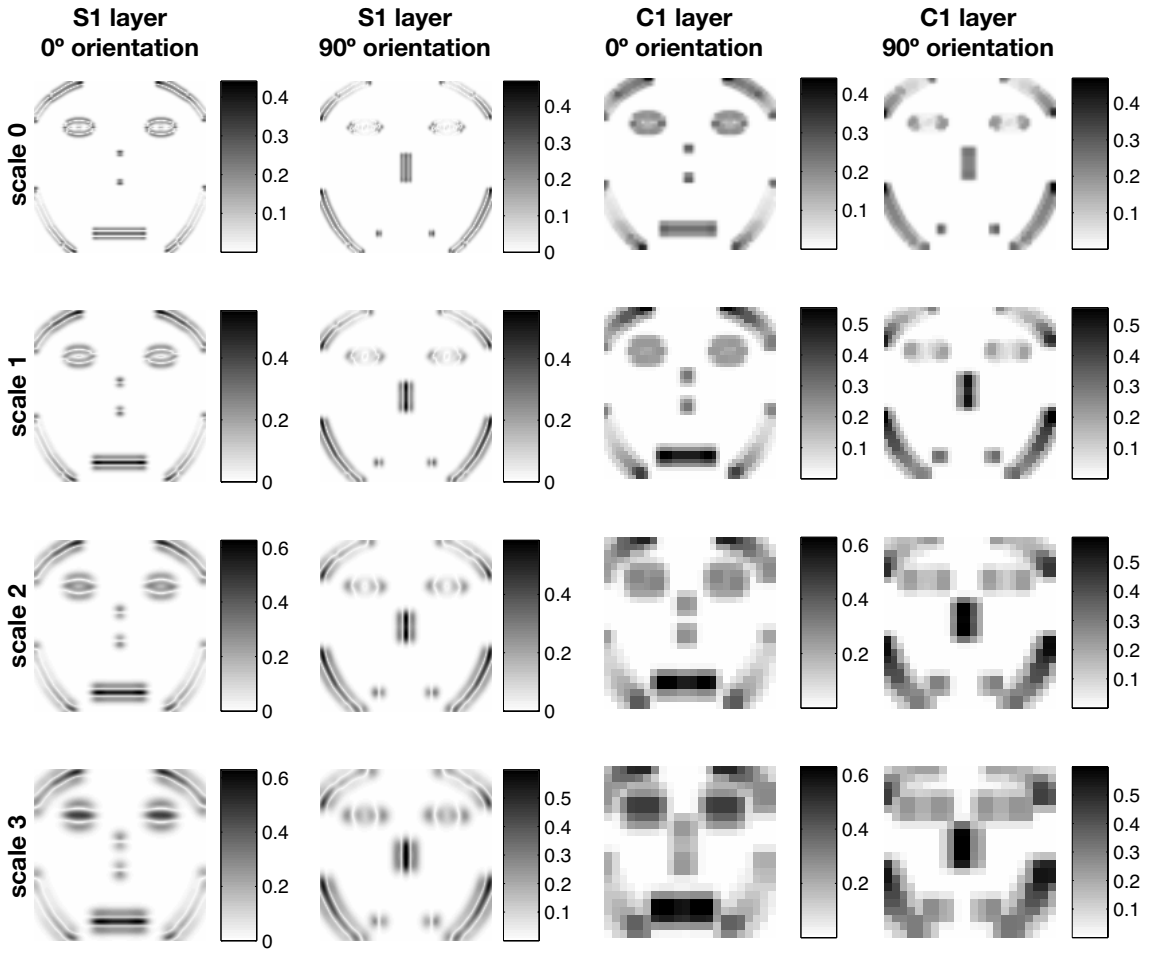


Figure 5.2. Responses from the S1 and C1 layers of a modified HMAX model operating on a single Brunswik face. Each image represents the responses of a set of units tuned to one of two orientations and one of four spatial scales; each column represents one of the possible orientations, and each row represents one of the possible spatial scales. The S1 responses are the immediate result of linear filtering, while the C1 responses are drawn from the S1 responses by pooling over spatial location and spatial scale using the *max* operator.

produced by pooling the activations of S1 units at neighboring spatial locations and across similar spatial scales. At the S2 level (Figure 5.3) more complex features are formed by pooling the activations of a 2×2 spatial array of neighboring C1 units tuned to specific orientations; in this way, different S2 units begin to represent features such as "elongated contour" or "corner" or "disk." Finally, each C2 unit pools across S2 units tuned to the same feature type, but at different spatial scales and/or spatial locations (Figure 5.4).

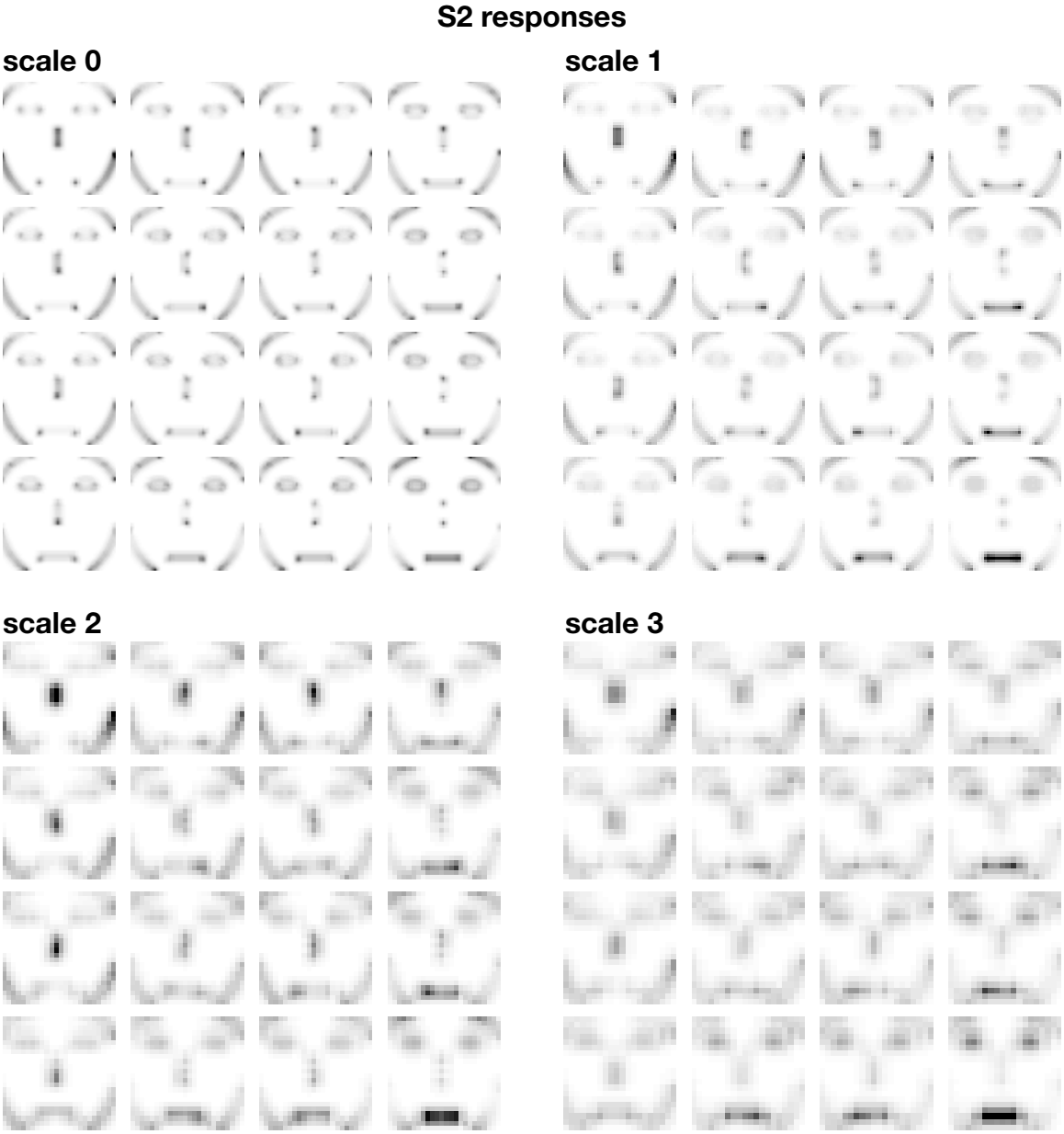


Figure 5.3. Responses from the S2 layer of a modified HMAX model operating on a single Brunswik face. Each group of images represents one of the four spatial scale bands; each of the 16 images within each group represents the responses of a batch of S2 units tuned to one of the 16 possible complex features (see Figure 5.5) formed by combinations of neighboring C1 units.

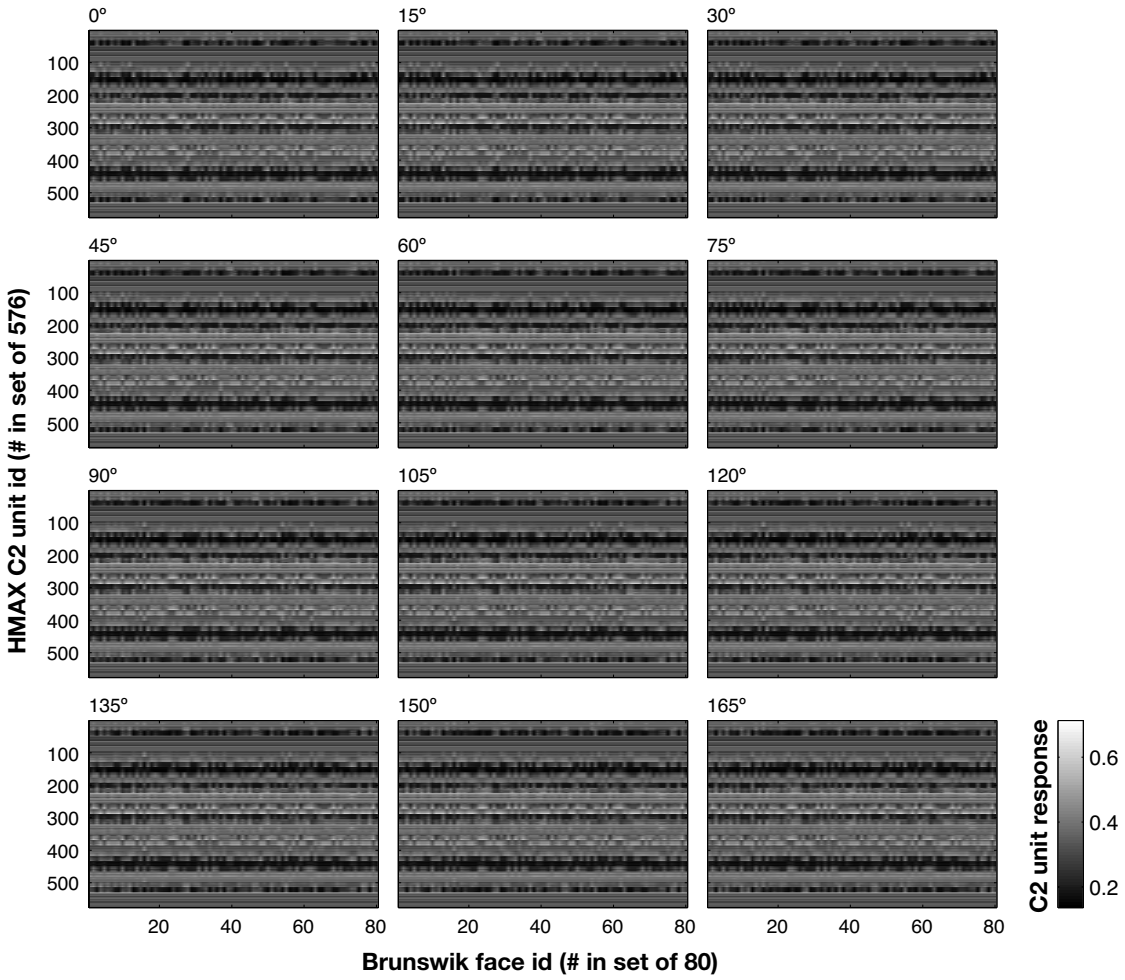


Figure 5.4. Each of the 12 subplots shown here represents one of the 80-member Brunswik face illustrated in Figure 3.7. Each point represents the response magnitude (given by grayscale value) of a C2 unit (numbered 1 to 576 along the y axes), for a given Brunswik face (numbered 1 to 80 along the x axes). Notice that the plots show strong horizontal "streaks," reflecting the fact that the responses of many of the C2 units are essentially invariant to the identity of the Brunswik face. As a result, the space of C2 units has a high redundancy, and can be efficiently transformed into a lower-dimensional representation as shown in Figures 5.8 and 5.9.

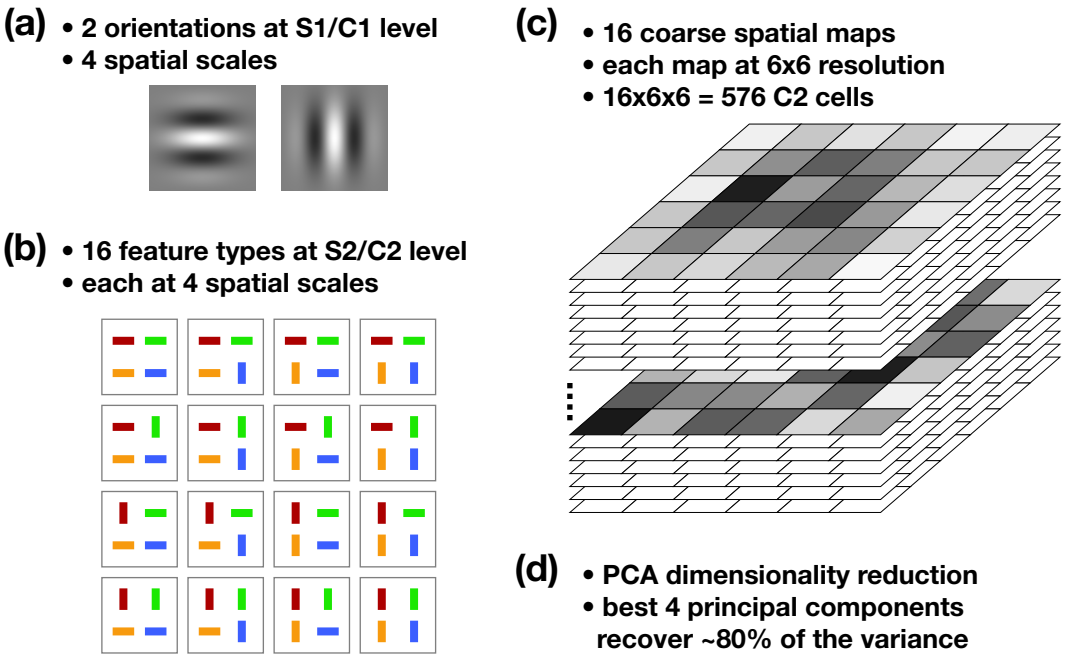


Figure 5.5. Shown here is a summary of the changes in our modified HMAX model, relative to the original model introduced by Riesenhuber and Poggio [1999] illustrated in Figure 5.1. **(a)** Only two orientations of linear filters, rather than four, were used at the S1/C1 level. These were found to be sufficient to represent the simple line-drawn stimuli under study. **(b)** With only two orientations, there were only 16 complex feature types at the S2/C2 level, rather than 256. **(c)** Rather than pooling over the entire image space, each C2 unit had a receptive field of 1/6 the length of the image, so that there were 36 C2 units for each of the 16 complex feature types. **(d)** In order to reduce the dimensionality of the output, principal component analysis (PCA) was applied to the 576 C2 units, and only the principal components corresponding to the 4 largest eigenvalues were selected for further use.

5.2 Modifications to HMAX

We made several modifications relative to the original model of Riesenhuber and Poggio [1999] (see Figure 5.5); these modifications were guided by the goal of increasing the variance of the HMAX outputs across the set of input images, so as to provide a rich but compact foundation for a subsequent categorization stage. First, instead of each C2 unit pooling across the entire image space, we subdivided the image into a 6×6 grid, with each C2 unit responding only to one of the 36 subregions. This increased granularity allowed the model to extract features that were more relevant as input to the categorization models. In addition, we restricted the number of orientation filters among the S1 units from four to two (*i.e.*, just hori-

zontal and vertical). This retained the model's ability to represent the variability among the simple schematic input images, but at the same time significantly reduced the dimensionality of the output space: since each S2/C2 feature type represented a four-part configuration of two possible S1/C1 orientations, there were $2^4 = 16$ S2/C2 feature types (rather than $4^4 = 256$ as in the original model). With 36 spatial locations, this gave a total of $36 \times 16 = 576$ C2 units.

5.3 C2 responses *vs.* the original representation

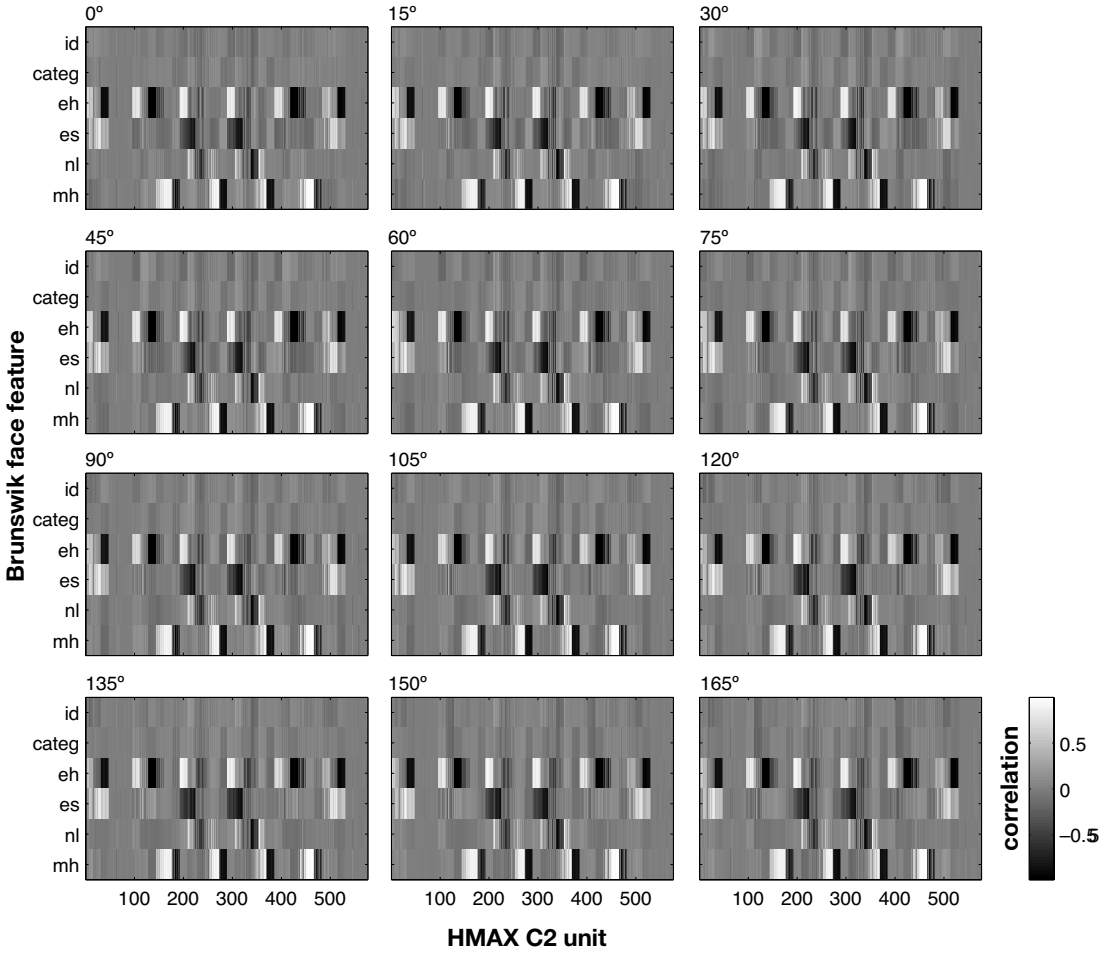


Figure 5.6. Each of the 12 subplots shown here represents one of the 80-member Brunswik face illustrated in Figure 3.7. Each point represents the correlation strength between one of the C2 units (numbered 1 to 576 along the x axes) with one of the parameters of the Brunswik faces (along the y axes). Abbreviations: *id*, identification number of the face within its containing set (*i.e.*, numbered 1 to 80); *categ*, the category identity of the face (*i.e.*, 1 or 2); *eh, es, nl, mh*: eye height, eye separation, nose length and mouth height.

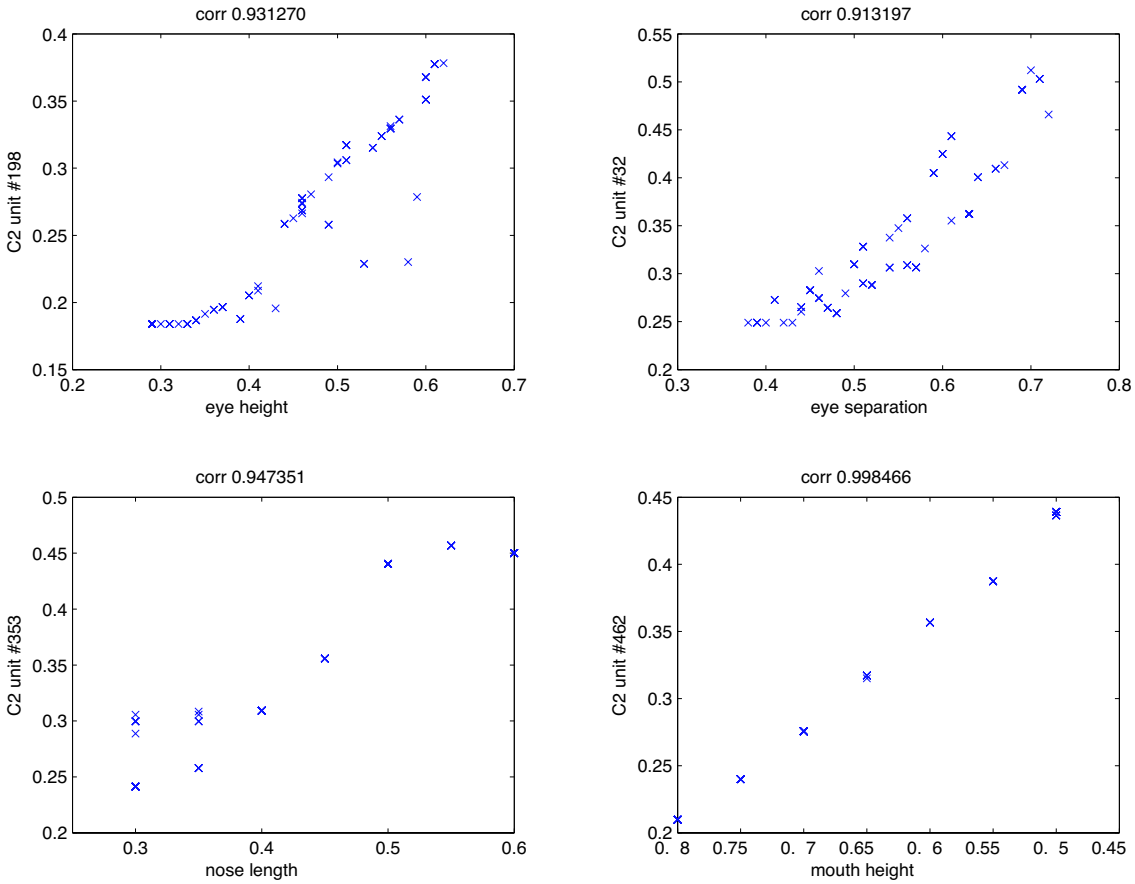


Figure 5.7. Shown here are the responses of several C2 units and the Brunswik face physical parameters between which there were strong correlations (*i.e.*, these are the units represented by bright white points in Figure 5.6).

Figure 5.6 shows how each of these 576 C2 units correlated with the features of interest in the Brunswik face sets from Figure 3.7. Many of the C2 units showed strong correlations with at least one of the original physical parameters of the Brunswik faces (EH: $\rho = 0.93$, ES: $\rho = 0.91$, NL: $\rho = 0.95$, MH: $\rho = 0.998$). The responses profiles of several such units are shown in detail in Figure 5.7. Notably, none of the C2 units showed a strong direct correlation with the category membership of the faces, emphasizing the need for a multi-stage categorization process in which a further categorization mechanism (like those described in Chapter 3) operates on an intermediate representation, perhaps similar to the one produced by the modified HMAX model.

In order to test how well the C2 output space of the modified HMAX model captured the variability in the input space of Brunswik faces, we performed the following test for each of the 12 stimulus configurations from Figure 3.7. First, we computed the variance in the output of each C2 unit across the set of the 80 in-

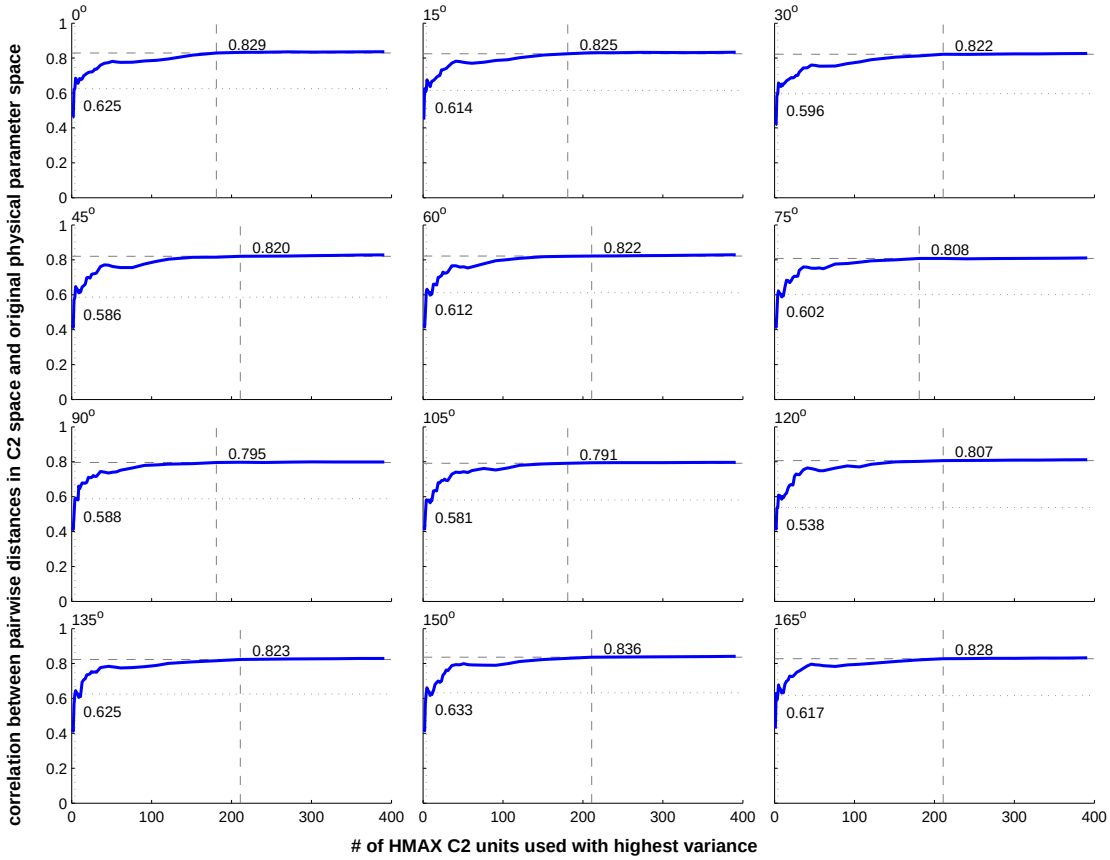


Figure 5.8. Each of the 12 subplots shown here represents one of the 80-member Brunswik face illustrated in Figure 3.7. Plotted here are the correlations between pairwise distances computed in the original parameter space and pairwise distances computed using the N C2 units with highest variance, where N increases along the x axes.

put stimuli, and sorted the C2 units from highest to lowest variance; presumably the units with highest variance are those that can best discriminate among the different input stimuli. Then, for varying N, we selected the N units with highest variance, forming an N-dimensional representation for each Brunswik face, and computed the pairwise Euclidean distances in this N-dimensional space between all possible pairs of faces. Finally, we computed the correlation across all pairs between these distances and the analogous distances obtained by using the original four-dimensional representation of the physical parameters (eye height, eye separation, nose length, mouth height). Figure 5.8 shows how these correlations vary as a function of N. Two numbers are of interest. First is the maximum possible correlation, regardless of N; this indicates how faithfully the full complement of C2 units capture the variability in the input space. This correlation was ≥ 0.8 for all of the 12 configurations. The second number of interest is how small of an N can

be used while still retaining an "acceptable" level of correlation between the two spaces. Since the original space was just four-dimensional, in principle an ideal intermediate representation would only need $N = 4$. In the case of the modified HMAX model, we found that the correlation values were ≈ 0.6 when only the first 4 C2 units with highest variance were selected.

5.4 PCA with C2 responses

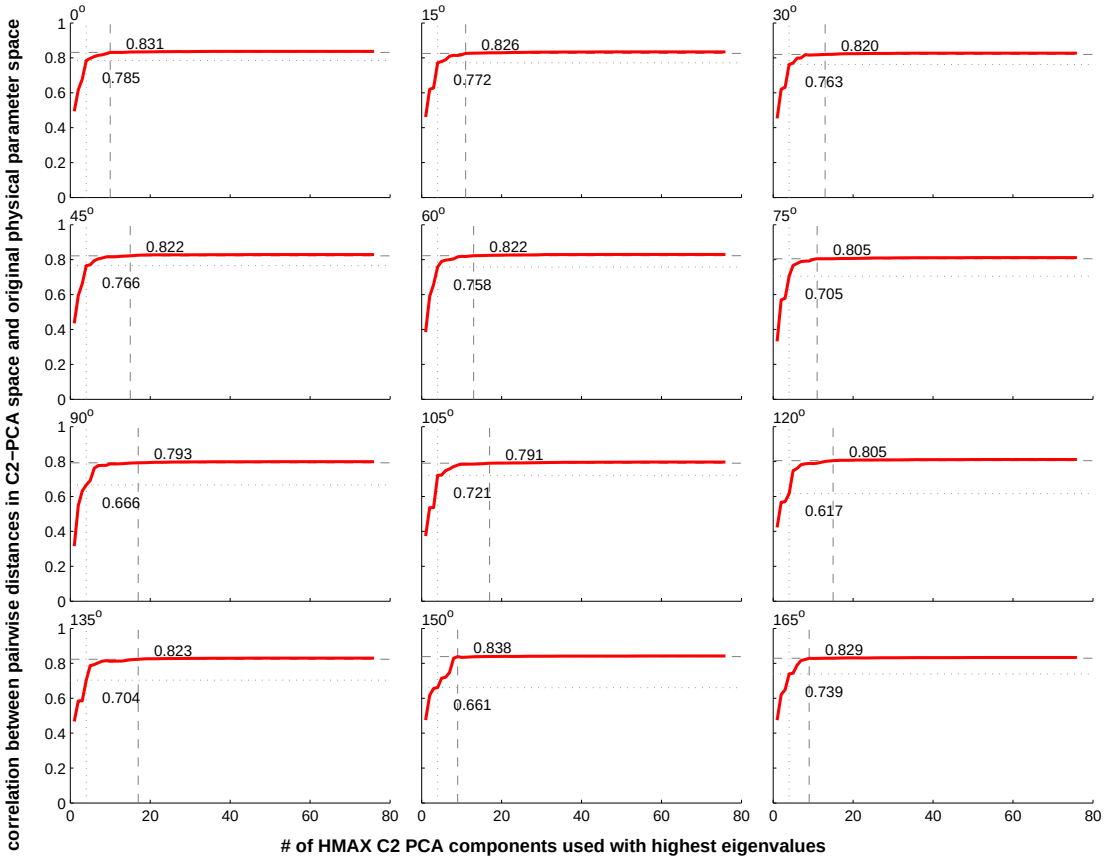


Figure 5.9. Each of the 12 subplots shown here represents one of the 80-member Brunswik face illustrated in Figure 3.7. Plotted here are the correlations between pairwise distances computed in the original parameter space and pairwise distances computed using the N PCA principal components (derived from C2 units) with highest eigenvalues, where N increases along the x axes. Note that the correlations here increase more quickly as a function of N than in Figure 5.8.

Since a significant amount of accuracy was lost by using only the first 4 of 576 C2 units with highest variance, we used principal component analysis (PCA) applied to the C2 units to retain the highest fidelity in as compact a representation

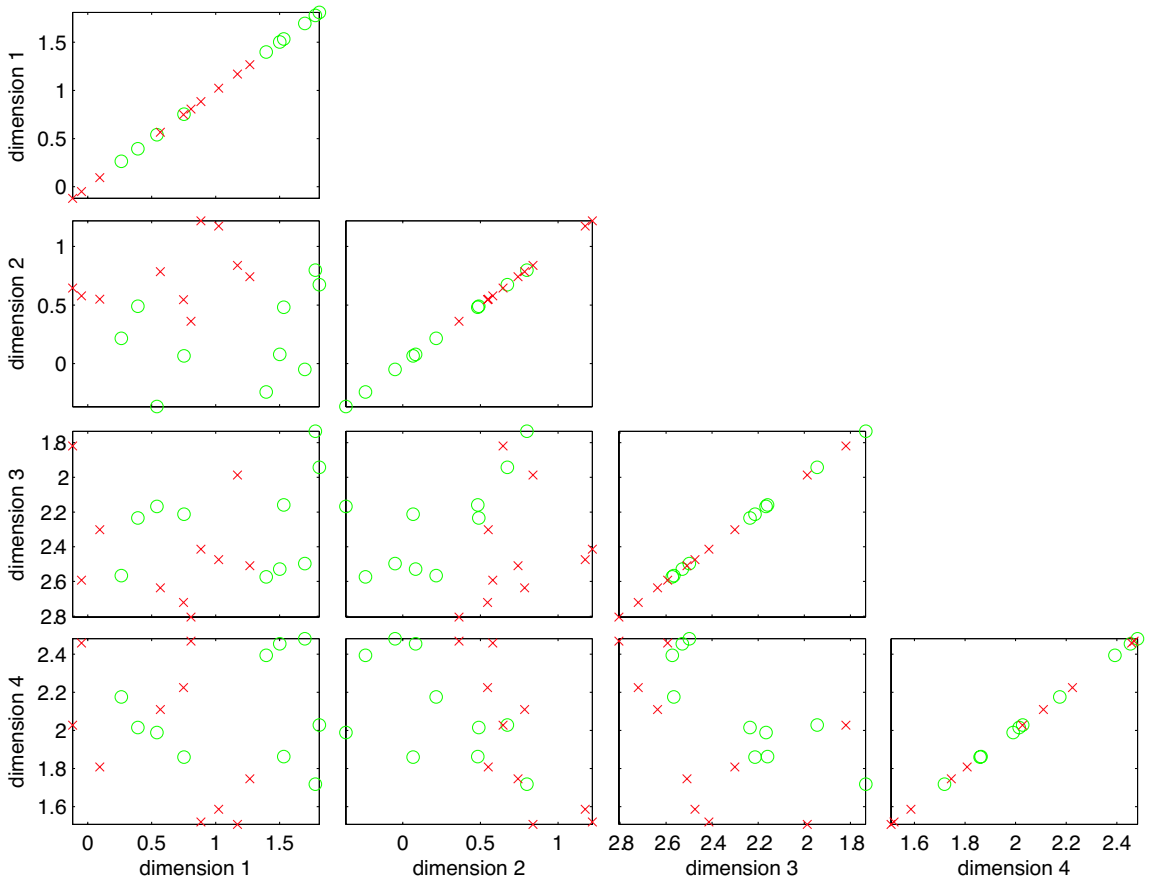


Figure 5.10. Shown here is the four-dimensional representation of one set of Brunswik faces obtained by selecting the first 4 principal components with highest eigenvalues derived with PCA from the C2 units. The two face categories are represented by $\times$ and $\circ$ in the plots. Note that the categories are nearly linearly separable in the projection onto dimensions 2 and 3.

as possible. Figure 5.9, like Figure 5.8, shows the correlations between pairwise distances as a function of the dimensionality of the representation, except that in this case the basis units of the representation are the principal components from PCA rather than the raw C2 units. Since PCA adds no new information, but only reshuffles existing information for optimal efficiency, we find that the maximum values in Figure 5.9 are the same as in Figure 5.8[1]: around 0.8 for each of the 12 sets. On the other hand we find that correlations approach this maximum much more quickly as a function of N, the number of principal components included. In general, we found that $> 95\%$ of the maximum correlation could be recovered with the first 50 of the 576 principal components, and $\approx 80\%$ of the maximum correlation was obtained with only the first 4 components. Therefore, for comparison with the

[1]Modulo rounding errors due to the limited accuracy of discrete floating-point arithmetic.

four-dimensional physical parameter configurations, we used the first 4 principal components from the modified-HMAX C2 activations to test how well the categorization models would fare with a biologically plausible input derived from the image space representation of the stimuli. Figure 5.10 shows the four-dimensional PCA configuration obtained for one of the 12 sets of Brunswik faces.

5.5 Categorization models using HMAX

Table 5.1. Goodness-of-fit of the models tested in Experiment 2. See also Table 3.3 for further discussion of the models' qualitative properties.

	RXM⟨1⟩	**RXM**⟨2⟩	**RXM**⟨3⟩	**SPC**⟨1⟩	**SPC**⟨2⟩	**SPC**⟨3⟩	**GCM**	**PBI**	**WPSM**
% var [orig]	89.36*	90.98*	91.49	89.36*	90.83*	**91.64**	86.84*	87.10*	84.90*
$-\ln L$ [orig]	75.72*	72.06*	71.32*	75.72*	71.65*	**69.92**	83.41*	83.66*	88.79*
AIC [orig]	**173.44**	178.13*	188.64*	**173.44**	177.30*	185.84*	178.81*	177.32	189.57*
% var [HMAX]	80.96*	83.98*	85.00*	80.96*	84.54*	**85.99**	75.62*	78.57*	72.57*
$-\ln L$ [HMAX]	91.24*	84.38*	81.89*	91.24*	82.77*	**78.11**	111.92*	97.70*	118.19*
AIC [HMAX]	204.48*	202.76*	209.78*	204.48*	**199.55**	202.23*	235.85*	205.40*	248.37*

% var, % of variance explained by model (larger value indicated better fit)
$-\ln L$, minus loglikelihood (smaller value indicates better fit)
AIC, Akaike Information Criterion (smaller value indicates better fit)
orig, models were fitted using objects represented by the original stimulus parameters, as in Chapter 3
HMAX, models were fitted using objects represented by features derived from a feed-forward early-vision network
bold numbers, model(s) which gave the best fit in each row
$*$, models whose fits were significantly worse ($p < 0.05$) than the best-fitting model in each row

Each of the models from Chapter 3 was re-fitted using the four-dimensional representations derived via PCA from the C2 activations of the HMAX model, and compared with models fitted using original physical parameters of the stimuli. Among the HMAX-based models, the SPC and RXM again gave better fits than the other models (see Table 5.1, rows 4–6). As before, the uncorrected measures (minus loglikelihood and %-variance) improved as the number of stored exemplars increased, with the best overall fit given by the SPC⟨3⟩. In contrast to the fits based on the physical parameters, the best AIC values were obtained with 2 (rather than 1) stored exemplars per category for both the SPC and RXM, although as before fits decreased again with more than 2 stored exemplars. Overall, the HMAX-based model fits were significantly poorer than the corresponding fits based on the physical parameters. Nevertheless, the absolute difference between the best-fitting HMAX-based and physical parameter-based models was only 5.6 %-variance.

With these results, we have begun to ground high-level models of categorization more firmly in neurobiology by combining them with an early vision model

(HMAX; Riesenhuber and Poggio, 1999) that encapsulates the processes that functionally precede object categorization in the visual system. Unlike the original categorization models which receive a high-level feature based description of their input, these hybrid models operate directly on a pixel-based image space representation of the input. Although the hybrid models fit relatively poorly when compared with the original models, their absolute performance is encouraging. The best-fitting HMAX-SPC$\langle 3 \rangle$ model was able to account for nearly 86% of the variance seen in subjects' responses. If anything, our results underestimate the capabilities of a hybrid model, since we used only the first 4 of 576 principal component vectors of the raw HMAX output, sacrificing $\approx$ 20% of the available variance. This performance was achieved using straightforward bottom-up processing of the input images, with no task-specific training or context-specific top-down modulation of the early vision stage. Yet, such top-down effects are certainly involved in the performance of human subjects, and the original high-level features are indeed a close approximation of subjects' internal representations as shown by MDS experiments. It thus appears that current high-level models of categorization can be linked to more detailed biological models of vision. A better integration of early-vision and object-categorization models—for example, by allowing attentional weights to propagate from the decision stage back to earlier sensory levels—is likely to uncover a more complete picture of the categorization process.

An open question is to what extent these computational insights, based on psychophysical experiments using simple, four-feature stimuli, carry over to the identification and categorization of complex objects in natural scenes. One challenge is to translate this analysis of the computational principles underlying object categorization into a mature understanding of how neurons along the ventral visual pathway can implement such operations [Sigala and Logothetis, 2002, Op de Beeck et al., 2001].

Part II

Attention

CHAPTER 6

Attention and eye movements

6.1 Introduction

Attention is the ubiquitous mechanism that regulates the bottleneck between the massively parallel world of sensation and the serial world of cognition [James, 1890]. This is particularly true in the visual system of primates, where 50% of the primary visual cortex is devoted to processing input from the central 2% (10°) of the visual field [Wandell, 1995]. In order to benefit from this non-uniform allocation of processing resources, the visual system relies on a combination of covert and overt attention-shifting mechanisms to efficiently bring behaviorally-relevant stimuli under the lens of central vision [Treue, 2003].

We used eye movements as an overt measure of where observers are directing their covert attention. This method is based on the pre-motor theory of attention [Rizzolatti et al., 1987], which suggests eye movements and attention shifts are driven by the same internal mechanisms. Links between eye movements and attention have been demonstrated by behavioral [Sheliga et al., 1994, 1995, Hoffman and Subramaniam, 1995, Kowler et al., 1995, Hafed and Clark, 2002] as well as physiological [Kustov and Robinson, 1996, Moore and Fallah, 2001, Moore et al., 2003, Moore and Fallah, 2004] and imaging [Nobre et al., 2000, Beauchamp et al., 2001] studies. A computational model of attention [Itti et al., 1998, Itti and Koch, 2001] has been shown to predict locations likely to be fixated by human observers with significantly better-than-chance accuracy [Parkhurst et al., 2002]. Despite these results, covert and overt attentional fixation locations may sometimes be

distinct [Posner and Cohen, 1984]; nevertheless it is likely that in the absence of explicit instructions to the contrary, overt and covert shifts of attention are closely related.

Our basic understanding of visual attention, as in all neuroscience, is primarily informed by direct sources such as measurement of brain activity (single-unit recording, EEG, MEG, fMRI) and of behavior (psychophysics). In the present study we used quantitative functional models to test hypotheses regarding the links between brain and behavior. Specifically, we asked whether, and to what extent, human fixation behavior is influenced by two putative physiological mechanisms reflecting the connectivity between neurons in V1. The first, covered in Chapter 7, corresponds to short-range interactions within a hypercolumn between units tuned having similar spatial receptive fields, but tuned to different orientations or spatial scales. The second, covered in Chapter 8, corresponds to long-range lateral connections between units having non-overlapping spatial receptive fields, and with orientation tunings that relate to the detection of elongated contours.

In both of these cases, we show that the mechanism being modeled has a significant influence on the locations that observers choose to fixate. More importantly, we find this effect in tasks involving free viewing of complex naturalistic scenes. Thus, the success of these models helps to support a quantitative link between observers' unconstrained overt behavior and the detailed functional properties of individual neurons as inferred from single-unit recordings and psychophysics experiments with constrained stimuli and task conditions. This detailed computational model of bottom-up, salience-based attention is useful for a range of applications from neuroscience to engineering. Machine vision systems face the same difficulties as do biological vision systems, and so a quantitative implementation of attentional selection can lead to similar improvements for machine vision systems. Indeed, models of bottom-up attention have been shown to improve the performance of traditional computer vision object recognition systems, both in training and in recall [Miau and Itti, 2001, Walther et al., 2002, Rutishauser et al., 2004]. Accurate models of behavior also serve a very practical goal in human-machine interface. Particularly for visual attention, there are many attention-demanding situations (*e.g.*, driving, flying) in which even a trained expert could occasionally benefit from an assistant system that was trained to match the expert's optimal behavior. None of this denies the crucial roles of top-down, task-dependent attention in conscious vision [James, 1890, Koch, 2004], yet in the absence of detailed quantitative models, we have concentrated here on the contribution of bottom-up,

salience-driven cues to fixation.

Twelve psychophysics subjects (ages 18–25) from the Caltech community participated as paid volunteers in the experiments described below. Informed consent was obtained from all subjects, and experimental procedures were approved by the California Institute of Technology's Committee for the Protection of Human Subjects.

6.2 Stimuli

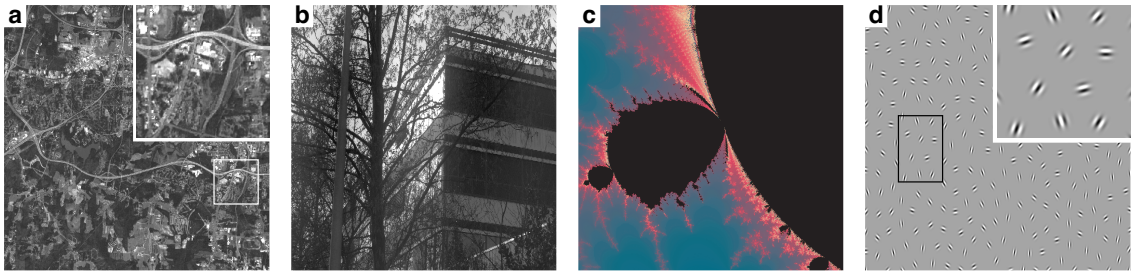


Figure 6.1. Samples from each of the image databases used for psychophysics and modeling experiments. All of the databases contained only grayscale images, except for the fractals which contained exclusively color images. Note that for didactic purposes, these same four exemplar images are used in subsequent figures to illustrate the output of each model component. **(a)** Overhead satellite imagery (grayscale; 10 m resolution "digital orthorectified"; publicly available from NGA). The inset provides a zoomed view of the boxed region. **(b)** Outdoor photographs (grayscale; Van Hateren database). **(c)** Computer-generated fractals (color; generated with *gnofract4d* software). **(d)** Gabor "snakes" and Gabor arrays — grayscale arrays of randomly spaced and oriented Gabor elements, some containing "snakes," or sequences of elements aligned so as to form a strong percept of a contour. The inset shows the boxed area at higher resolution. Although the "snake" is not highly visible at the scale shown here, these contours are strongly salient when viewed at the scale used in our psychophysics experiments.

We used four classes of images (Figure 6.1), ranging in size from 1000×1000 to 1536×1024 pixels, for a visual angle subtended of roughly $15.8^\circ \times 15.8^\circ$ to $16.2^\circ \times 25^\circ$. The experiments reported here typically included roughly 100 images from each image class:

- grayscale *overhead satellite imagery*, 10-meter resolution "digital orthorectified" (DOI10m) [1];
- grayscale *outdoor photographs*[2] [van Hateren and van der Schaaf, 1998];

[1] From the National Geospatial-Intelligence Agency (NGA) (`http://geoengine.nga.mil/`)
[2] `Available from http://hlab.phys.rug.nl/imlib/`

- computer generated color *fractals*[3];

- grayscale *Gabor "snakes"* and *Gabor arrays* containing arrays of Gabor elements with random orientations, phases, and spatial locations, generated with a previously-described algorithm [Braun, 1999a,b]; some of the arrays included "snakes," or sequences of Gabor elements with orientations aligned so as to induce a strong percept of a contour, even though element spacing and Gabor phase were otherwise random (see also Figure 8.1 for a more detailed view of these Gabor arrays).

6.3 Free-viewing task

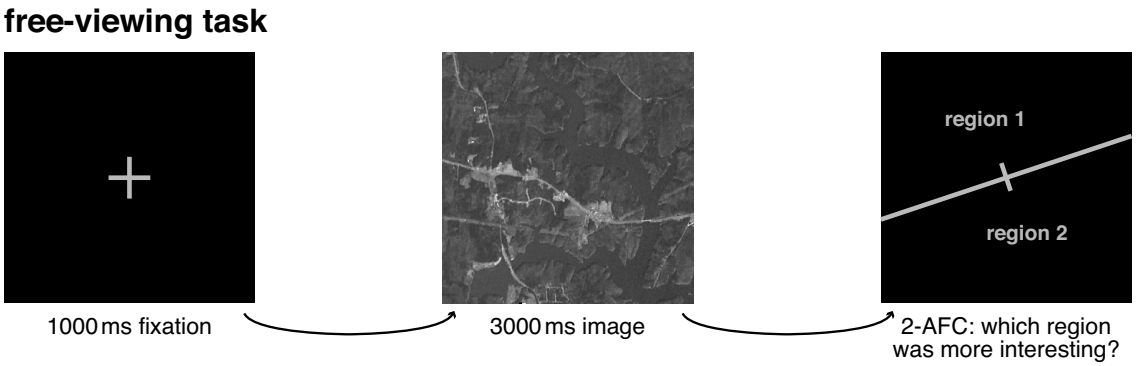


Figure 6.2. Illustration of the free-viewing task that subjects performed while their eye movements were recorded with an infrared eye tracker operating at 120 Hz. Each trial began with a fixation cross (1000 ms), followed by a stimulus image (3000 ms) drawn from one of the image categories shown in Figure 6.1. After the image disappeared, a subjects were presented with a single line bisecting the screen into two regions, and were asked to make a two-alternative forced choice (2-AFC) as to whether they thought "the most interesting point" in the just-seen image fell in region 1 or region 2. The orientation of the line varied from trial to trial; since subjects could not predict the orientation, they were forced to consider the entire stimulus image, without being encouraged to focus on any particular aspect of the image.

Images were presented to subjects in a free-viewing task (Figure 6.2). Each trial began with a 1000 ms fixation cross at the center of a blank screen, which subjects were instructed to fixate. This imposed some consistency on the initial conditions of the subsequent scanpaths, across different images and observers. Following the

[3]Some of the fractal images were drawn from the online Spanky Fractal Database (http://spanky.triumf.ca/) and others were custom-designed with freely-available *gnofract4d* software (http://gnofract4d.sourceforge.net/).

fixation cross, a target image was shown for 3000 ms. Subjects were instructed to "look around the image" with no restrictions except the knowledge that they would have to provide a response, as follows. Immediately after the target image disappeared, a single line was presented at an arbitrary orientation bisecting the screen into two regions of equal size. The two regions were labeled as "1" and "2," and subjects were required to make a button press indicating which region contained the location that they had found "most interesting" in the previous image. Our motivations for requiring this response were twofold:

- to encourage subjects to be vigilant in their task and engage in active eye movements (without a minimal task to motivate them, subjects might efficiently choose to make no eye movements at all); and
- to avoid imposing any particular top-down bias on the task (such as would occur if subjects were asked to search for horizontal lines, or to judge the brightness of the image, or to name objects in the image), allowing direct comparisons with a model of bottom-up attention.

Although no time limit was imposed on the responses, subjects were encouraged prior to the experiment not to dwell on the choice for too long, but rather to make their best guess if they felt unsure.

6.4 Eye tracking

Subjects were seated 75 cm from a CRT used for stimulus display, which covered $26^\circ \times 19^\circ$ of visual angle, and were asked to use a chinrest in order to minimize eye-tracking errors due to head movements. We used an infrared (IR) eye-tracking system (ISCAN, Inc.) to sample and record subjects' eye position at 120 Hz (see Figure 6.3). An illuminator and camera were placed ~65 cm from the subject, and his or her right eye was illuminated with a beam of low-intensity (~1 mW/cm^2) invisible IR light (~850 nm). The camera recorded a closeup image of the eye, which was processed in real time to extract the positions of two features: (1) $\boldsymbol{p}$, the IR-dark spot at the center of the pupil, and (2) $\boldsymbol{c}$, the IR-bright spot where the IR beam produces a specularity on the cornea. The vector difference $\boldsymbol{v'} = \boldsymbol{p} - \boldsymbol{c}$ of these two positions gives a measure of eye position that is independent of head position. An empirical correspondence between $\boldsymbol{v'}$ (in camera coordinates) and the subject's real-world point-of-regard $\boldsymbol{v}$ (in stimulus display coordinates) was

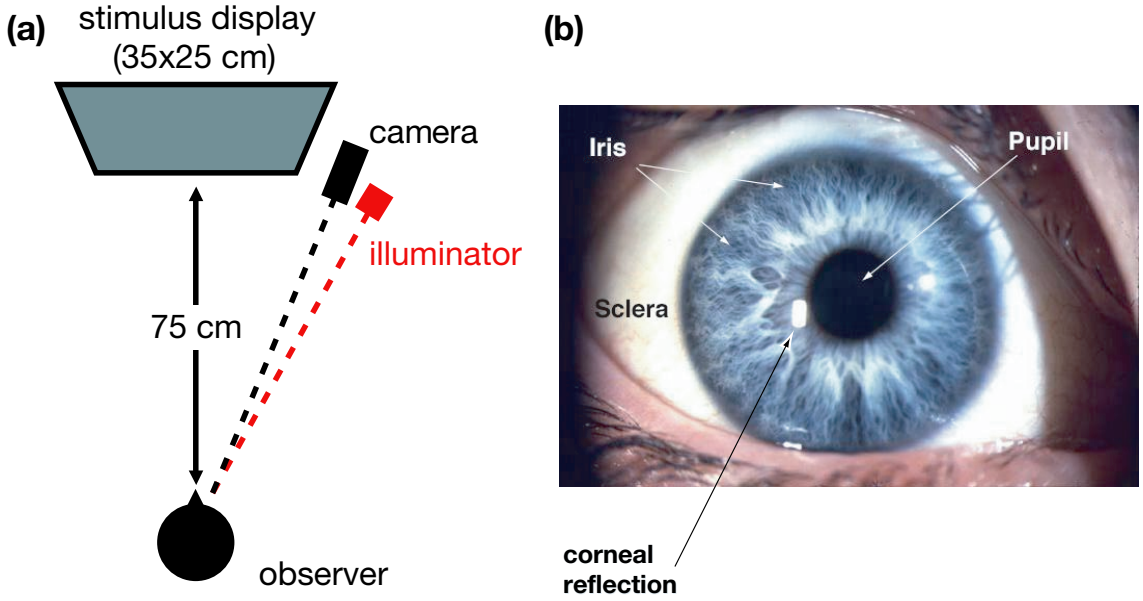


Figure 6.3. **(a)** Shown here is a diagram of the setup used in eye-tracking experiments. Each subject was seated with his or her chin in a chinrest, 75cm from a computer monitor on which stimulus images were displayed. The right eye was illuminated with invisible infrared light, and was recorded with an infrared camera. The camera image was similar to that shown in **(b)**; the features of interest for inferring subjects' eye position were the center of the pupil and the center of the corneal reflection generated by the infrared light beam. Under the assumption that the eye is roughly spherical, the corneal reflection does not move during a pure eye movement (the eye rotates within the socket), and during a translational head movement the pupil does not move relative to the corneal reflection.

established by a set of calibration trials in which the subject fixated a series of crosses shown at 25 different locations on an invisible 5×5 grid in the stimulus display (see Figure 6.4 for an illustration of this process). These v-v' pairings could then be used to interpolate the subject's point-of-regard throughout the remainder of the session. Following every session, each lasting about 12 minutes, we re-recorded subjects' eye positions at the 25 calibration locations in order to assess how much drift had occurred during the recording session. Across 300 calibration points (4 subjects, 3 sessions per subject, 25 points per session), the mean error was 0.37° degrees of visual angle.

For subsequent comparison with salience models, we used the human eye-tracking data both in its raw form (a 120 Hz time-series of spatial locations) as well as in a processed form in which we parsed each scanpath into a series of fixation intervals. This process is illustrated and described in Figure 6.5.

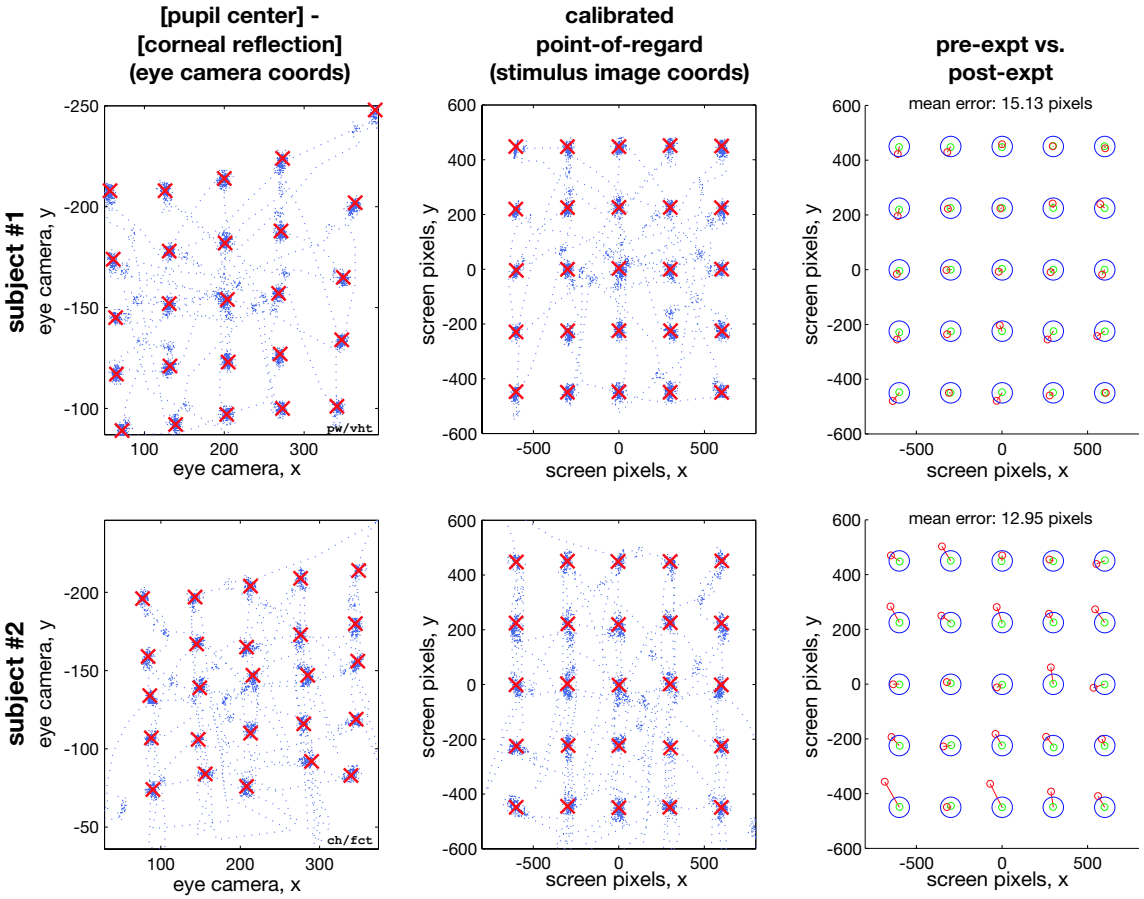


Figure 6.4. These plots illustrate the calibration process for two example subjects (row 1 and row 2). Subjects initially fixate 25 points on a 5×5 grid (clusters in left column), and the position of the pupil center relative to the corneal reflection is recorded. Since the corresponding screen coordinates of these points are known, an empirical mapping relation can be established between eye position in camera coordinates, and calibrated point-of-regard in screen coordinates. The center column shows the same fixation sequence at left, but transformed into screen coordinates. Finally, subjects were tested on the same 25-point fixation grid following each experiment, and these data were mapped into screen coordinates using the mapping defined by the pre-experiment fixation sequence. The right column shows the calibrated screen coordinates from the pre-experiment (green central circles) and post-experiment (red circles offset from grid); any differences reflect calibration errors likely due to rotational head movements by the subject or to changes in the wetness (and hence the reflectivity) of the retina over the course of the experiment,

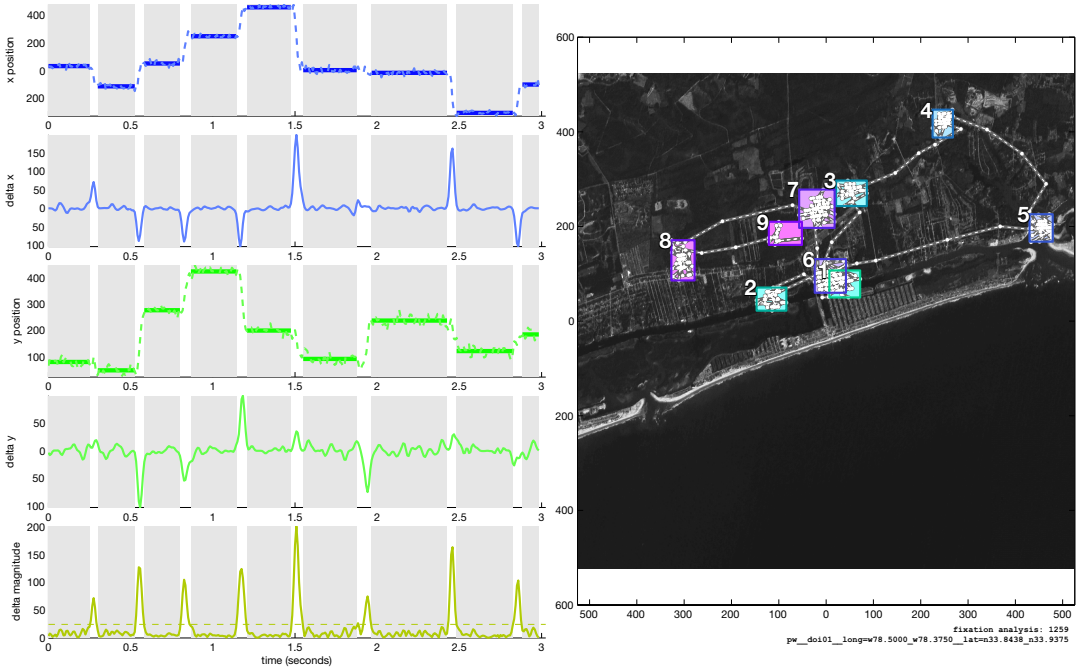


Figure 6.5. Shown here is an illustration of the method used to parse a scanpath into a sequence of fixation intervals. The raw data were 120 Hz sequences of x and y position in screen coordinates, following the calibration method outlined in Figure 6.4. These were first smoothed slightly with a 5-point median filter; sample results are shown in the first (x) and third (y) traces from the top at left. Second, the x and y velocity components were estimated by convolving the smoothed position traces with a $[-1\ 0\ 1]$ filter; results are shown in the second and fourth traces. Finally these were used to compute the magnitude of the velocity (bottom trace), and an adaptive cutoff value was chosen (dashed line) to differentiate between steady fixation intervals and saccade or eyeblink intervals. The intervals identified as fixations are indicated by the gray boxes beneath each of the traces at left and by the solid lines within the position traces. At right, the original scanpath is shown overlaid on the stimulus image, and the fixation intervals are marked by boxes whose numbers indicate the temporal order.

6.5 Salience model

All of the models described here[4] are based on the computational architecture of a salience model of bottom-up visual attention first proposed by Koch and Ullman [1985] and developed in detail by Itti et al. [1998] (see Figure 6.6). Each input image is processed in parallel through a number of feature channels (*e.g.*, one each for color, luminance, orientation), and the outputs of these channels are ultimately combined to form a single salience map. This map ascribes a scalar value to each

[4]Source code for the iLab Neuromorphic Vision Toolkit (iNVT), including the salience model and each of the extensions described below, is freely available under the GNU General Public License (GPL) at `http://ilab.usc.edu/toolkit/`.

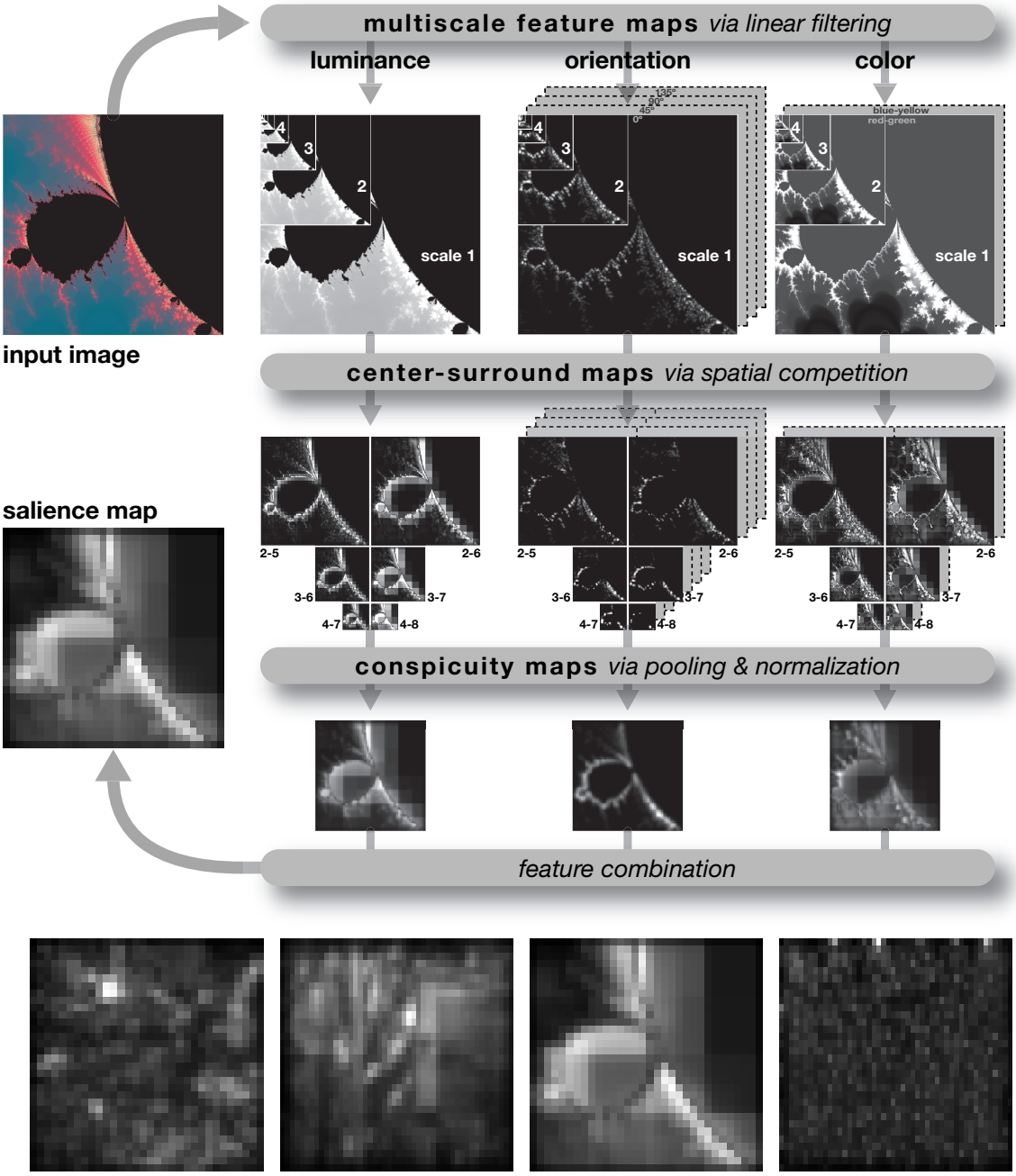


Figure 6.6. Schematic diagram of the salience model (top) and salience maps corresponding to the four exemplar images from Figure 6.1 (bottom row). In the salience model, an input image is processed in parallel through multiple channels. In each channel (here for luminance, orientation, or color), the image is filtered at nine spatial scales, and the resulting feature maps pass through a center-surround operation to accentuate contrast. The center-surround maps are combined across spatial scales leading to one conspicuity map per channel, and finally these conspicuity maps are combined across features to produce a single feature-independent salience map. Additional channels may be included in parallel to the three channels shown here; in our experiments, we tested a modified orientation channel that included short-range orientation interactions (Figure 7.1) and a contour-integration channel based on long-range orientation interactions (Figure 8.2).

point in the input image, indicating how salient or "interesting" that location is, regardless of which features contributed to the salience.

The individual channels share a common architecture as well. In general, the input image is first passed through a series of linear filters at nine spatial scales to form a dyadic pyramid. These filter outputs are then subject to spatial competition via a center-surround operation, implemented as a difference between fine and coarse scales in the pyramid. Typically there are six *feature maps* generated by this center-surround operation, using center scales $c \in \{2, 3, 4\}$ and surround scales at $s = c + \delta$, with $\delta \in \{3, 4\}$. The feature maps are summed across scales and passed through a nonlinear normalization operation designed to reduce or eliminate numerous weak local maxima in favor of a small number of stronger near-global maxima. This produces a single *conspicuity map* representing the output of the channel; these conspicuity maps are eventually summed across channels and renormalized to produce the final salience map.

The standard channels for static images include an luminance channel that responds to luminance contrast, an orientation channel (including filter outputs from multiple scales and orientations) that responds to orientation contrast, and a color channel that responds to opponent-color contrast. These reflect many of the fundamental computational operations thought to be performed in the early stages of the visual system [Marr, 1982, Wandell, 1995]. Nevertheless, the modular architecture of the salience model allows other new channels to be included in parallel to the standard channels, or even to replace one or more of them. This is the approach we used in testing more detailed models of interactions among orientation-tuned units, as described next.

6.6 Comparing model and eye-tracking data

Our analyses rested on the degree of correspondence between human fixation locations and model salience maps. The most straightforward approach was as follows (Figure 6.7). Each salience map was linearly normalized to have zero mean and unit standard deviation. Next, the normalized salience values were extracted from each point corresponding to the fixation locations along a subject's scanpath, and the mean of these values, or *normalized scanpath salience* (NSS), was taken as a measure of the correspondence between the salience map and scanpath. Due to the pre-normalization of the salience map, normalized scanpath salience values greater than zero suggest a greater correspondence than would be expected by

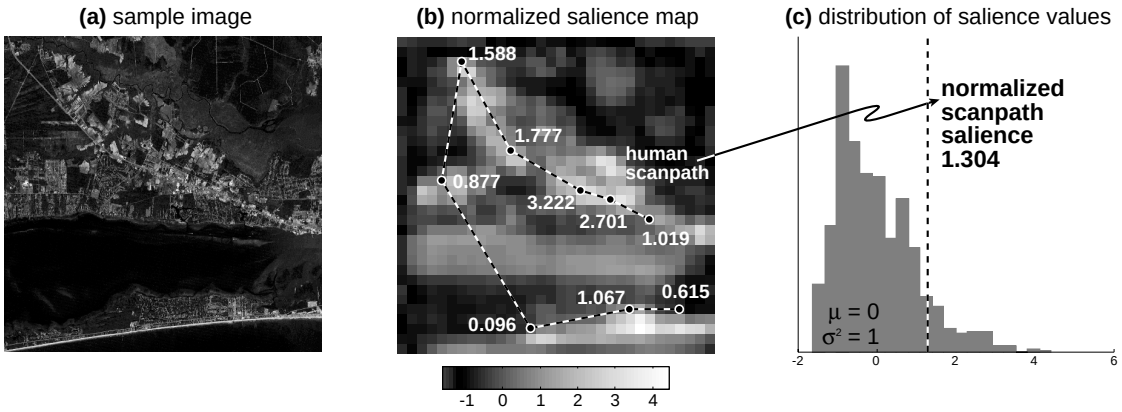


Figure 6.7. Illustration of the method used to compare fixation locations obtained from eye tracking with salience maps obtained from various computational models. **(a)** A sample image is shown to both the human observer and the model. **(b)** The model generates a salience map (grayscale image), which is normalized to have zero-mean and unit standard deviation (see scalebar). A series of fixation locations is generated by the observer (connected dots), and the normalized salience value is extracted for each location (values are shown here next to the corresponding fixation locations). **(c)** The average normalized salience value across all fixation locations is taken as the *normalized scanpath salience* (NSS), and compared against the distribution of salience values across the entire salience map (gray histogram). For the scanpath shown here, the normalized scanpath salience indicates that, on average, the model-predicted salience at fixated locations was 1.304 standard deviations above chance level. Since the NSS is scale-free, it can be used to compare the degree of correspondence between observed and predicted behavior for different observers and images.

chance between fixation locations and the salient points predicted by the model; a value of zero indicates no such correspondence, while values less than zero indicate an anti-correspondence between fixation locations and model-predicted salient points. Also because of the pre-normalization, these measures could be compared across different subjects, image classes, and model variants; with such a data pool, statistical tests indicated whether the distribution of NSS values was different from the zero-mean distribution that would be expected by chance.

Our approach is similar to that taken by Parkhurst et al. [2002] in that both rely on a linear transformation of salience values; however, our approach uses a variable dynamic range based on the variance of the salience values, while the alternate approach uses a fixed dynamic range based on the difference between the minimum and maximum values (which were rescaled to 0 and 100, respectively, in Parkhurst et al., 2002). In addition, our approach compares salience values at fixated locations to chance distributions unique to each image; the alternate approach compares salience values to a single chance distribution based on all images in a

Table 6.1. Shown here is a summary of the fits between each model and the scanpaths recorded during the free-viewing task (Figure 6.2). Each number represents the average *normalized scanpath salience* (NSS) value, for a given model, across all of the fixation locations recorded while observers freely viewed images for 3000 ms each. The NSS values were obtained by the method illustrated in Figure 6.7, in which salience maps were first normalized to have zero mean and unit standard deviation, and then for each scanpath the average normalized salience was computed for the fixation locations along the scanpath. Thus for the data shown here, a value of zero would indicate the absence of a correspondence between model predictions and observed fixation locations; a value of one would indicate that, on average, the model-predicted salience was one standard deviation above chance at each fixation location for all observers and all images in the given image category. The first three rows show these correspondences for salience maps predicted by a random model, the baseline salience model (Figure 6.6), and the control condition in which the "salience map" is derived from all observers' scanpaths. This last condition quantifies how well the pooled fixation locations from all observers predict the specific fixation locations of individual observers; as such, it provides a theoretical upper limit for the performance of the models, since the models are not designed to account for inter-observer variability. Thus, the next three rows express the performance of each model as a percentage of the corresponding upper limit.

	Outdoor	**Fractal**	**Satellite**	**Gabor snake**	**Gabor array**
NSS					
Random model	-0.01	-0.02	0.02	-0.01	0.02
Baseline salience model	0.69	0.44	0.62	0.10	0.14
Inter-observer	1.30	1.13	1.10	1.15	0.91

given image category. Our method was intended to accommodate the wide variety of salience distributions observed for different input images (for example, consider how a salience map with 100 points, 90 with value 1.0 and 10 with value 0.0, would be handled relative to a second salience map with 100 values spaced evenly between 0.0 and 1.0).

A summary of all of the fits between models and human behavior in the free-viewing task is given in Table 6.1. Each number gives the average NSS across all observers and images in that image class. In general, our data agree with previous results [Parkhurst et al., 2002] showing that the baseline salience model was significantly above chance ($p < 10^{-23}$) at predicting locations likely to be fixated by observers in a free-viewing task. As expected, this result was largely independent of image category for naturalistic images such as the overhead imagery, outdoor photos, and fractals, but did not hold for more artificial images such as the Gabor arrays, for which the baseline salience model was virtually at chance in predicting fixation locations. Indeed, we chose to use the Gabor arrays for exactly this reason:

Table 6.2. Shown here is the performance of each model (average normalized scanpath salience, NSS) as a percentage of the corresponding upper limit given by the NSS of the inter-observer model. The arrangement of the table is the same as in Table 6.1.

	Outdoor	Fractal	Satellite	Gabor snake	Gabor array
NSS % of Inter-observer NSS					
Random model	0%	-2%	2%	-1%	2%
Baseline salience model	53%	39%	57%	9%	15%
Inter-observer	100%	100%	100%	100%	100%

nothing in the baseline model can "see" the contours, yet they are perceptually salient to observers.

The theoretical range of NSS values is bounded from below by the behavior of a random "model," in which the salience maps simply contain noise drawn from a normal distribution, and bounded from above by the behavior of an inter-observer "model" in which the salience maps are generated by the pooled fixation locations from all observers. The very nature of our analysis method requires that the random model should produce NSS values of 0, and indeed we find values that are nearly 0 (slight differences from 0 are due to the finite size of our data set). On the other end, we find that the inter-observer yields NSS values between 0.9 and 1.3, depending on the image type. So one way to intuitively understand the performance of the salience model is to consider its performance as a percentage of the difference between the fit of the random and inter-observer "models." These values are shown in Table 6.2 and range from 39% to 57% for the natural image classes and from 9% to 15% for the Gabor arrays.

6.7 Discussion

Our experiments were based on a simple attempt to explain human behavior in an image-viewing task with purely bottom-up models of attention; we have disregarded important top-down contributions from attentional state, past experience, and inter-observer differences, in order to assess how much can be predicted from bottom-up influences alone. In this respect our method follows that of Parkhurst et al. [2002], and our results with the baseline model are in agreement as well: we found highly significant correspondences between model predictions and human fixation locations. However, the main focus of the present study was to extend this method to test, via more specific models, whether certain early vision mechanisms

play a significant role in determining subjects' fixation locations, a task that we will undertake in Chapters 7 and 8.

We have relied on an assumption of a substantial overlap between the biological mechanisms responsible for covert attention shifts and overt eye movements; on this "pre-motor theory of attention" [Rizzolatti et al., 1987], attention shifts are essentially planned saccades whose motor execution is inhibited. This is supported by behavioral evidence showing that, despite motor inhibition, the spatial locus of attention exerts a small but detectable influence on the trajectories of subsequent saccades [Sheliga et al., 1994, 1995, Kustov and Robinson, 1996] or on the distribution of microsaccades during fixation [Hafed and Clark, 2002]. These results suggest that computational models of attention and saccadic eye movements would be nearly identical up until the execution stage, where the dynamics would be expected to change due to the motor inertia of eye movements or the differing strengths of inhibition-of-return. Indeed, it is plausible that other modes of behavioral output, such as verbal report or finger-pointing, could be driven by the same core mechanisms. Chapter 9 will describe ongoing work using such approaches to further explore which computational elements are intrinsic to spatial attention, and which are specific to particular output modalities [Briand et al., 2000, Astafiev et al., 2003].

In working with the salience model, we found that not only was there a strong correspondence between model and fixation locations, but that in addition the model accounted for a large overall fraction of the observed behavior. Allowing that this general model of vision is not intended to account for inter-observer differences, an absolute upper limit on the performance of such models is given by the ability to predict one subjects' behavior from the average behavior of the remaining subjects. As shown in Table 6.2, the performance of the models we tested was around 50% of this theoretical limit; as a crude measure, this suggests that the models accounted for roughly half of the variance in spatial positions of fixated locations, outside of inter-observer differences. In the next chapters, we will examine whether more detailed models of specific visual processing mechanisms can bring us closer to this theoretical inter-observer upper limit.

CHAPTER 7

Short-range orientation interactions

7.1 Introduction

Short-range interactions among orientation-tuned units in V1 with retinotopically overlapping receptive fields have been shown to play an important role in detection and discrimination thresholds in a range of psychophysical tasks involving Gabor-like grating stimuli [Lee et al., 1999]. Such interactions can be implemented as a form of divisive inhibition leading to contrast-enhancement—similar to a center-surround operation, but operating in the orientation and frequency domain rather than in the spatial domain. Furthermore, divisive inhibition provides the gain control needed to work within the limited dynamic range of neurons [Heeger, 1992]. In this chapter, we ask whether a salience model that includes short-range orientation interactions could account for behavior in a naturalistic free-viewing task that would not have been explained by the baseline salience model.

7.2 A model of short-range orientation interactions

We adapted a model of interactions among overlapping orientation-tuned units [Lee et al., 1999, Itti et al., 2000] (see Figure 7.1) that could be substituted for the standard orientation channel in the salience model. In this enhanced orientation channel, orientation-sensitive units tuned to overlapping spatial locations, but to different orientations θ and spatial frequencies ω, form an inhibitory pool. In

the two-stage model, the feed-forward first-stage response $E_{\theta,\omega}$ is subject to self-excitation and suppression from the inhibitory pool. The result of these interactions is the non-linear second-stage response $R_{\theta,\omega}$, given by

$$\begin{aligned} R_{\theta,\omega} &= \frac{(E_{\theta,\omega})^{\gamma}}{S_{\delta} + \sum_{\theta',\omega'} W_{\theta\theta',\omega\omega'} (E_{\theta',\omega'})^{\delta}} \\ \delta, \gamma &: \text{power-law exponents} \\ S &: \text{semi-saturation constant} \\ W_{\theta\theta',\omega\omega'} &= e^{-\frac{(\theta-\theta')^2}{2\Sigma_{\theta}^2}} e^{-\frac{(\omega-\omega')^2}{2\Sigma_{\omega}^2}} \\ \Sigma_{\theta}, \Sigma_{\omega} &: \text{widths of inhibitory pool} \end{aligned}$$

It should be noted that in the original model, the feed-forward responses $E_{\theta,\omega}$ were calculated using ideal filters tuned for a given θ and ω:

$$\begin{aligned} E_{\theta,\omega} &= A c_s e^{-\frac{(\theta_s-\theta)^2}{2\sigma_{\theta}^2}} e^{-\frac{(\omega_s-\omega)^2}{2\sigma_{\omega}^2}} + B \\ c_s &: \text{stimulus contrast} \\ \theta_s &: \text{stimulus orientation} \\ \omega_s &: \text{stimulus spatial frequency} \\ \sigma_{\theta} &: \text{sharpness of orientation tuning} \\ \sigma_{\omega} &: \text{sharpness of spatial frequency tuning} \\ A &: \text{contrast gain} \\ B &: \text{background activity level} \end{aligned}$$

In contrast, for the modified version that was incorporated into the salience model, we simply used the values already computed in the dyadic orientation-tuned pyramids. Lee et al. [1999] used an extensive series of psychophysics experiments to calibrate the interactions in this model, and we used the same calibrated values in the version of that was included into the salience model.

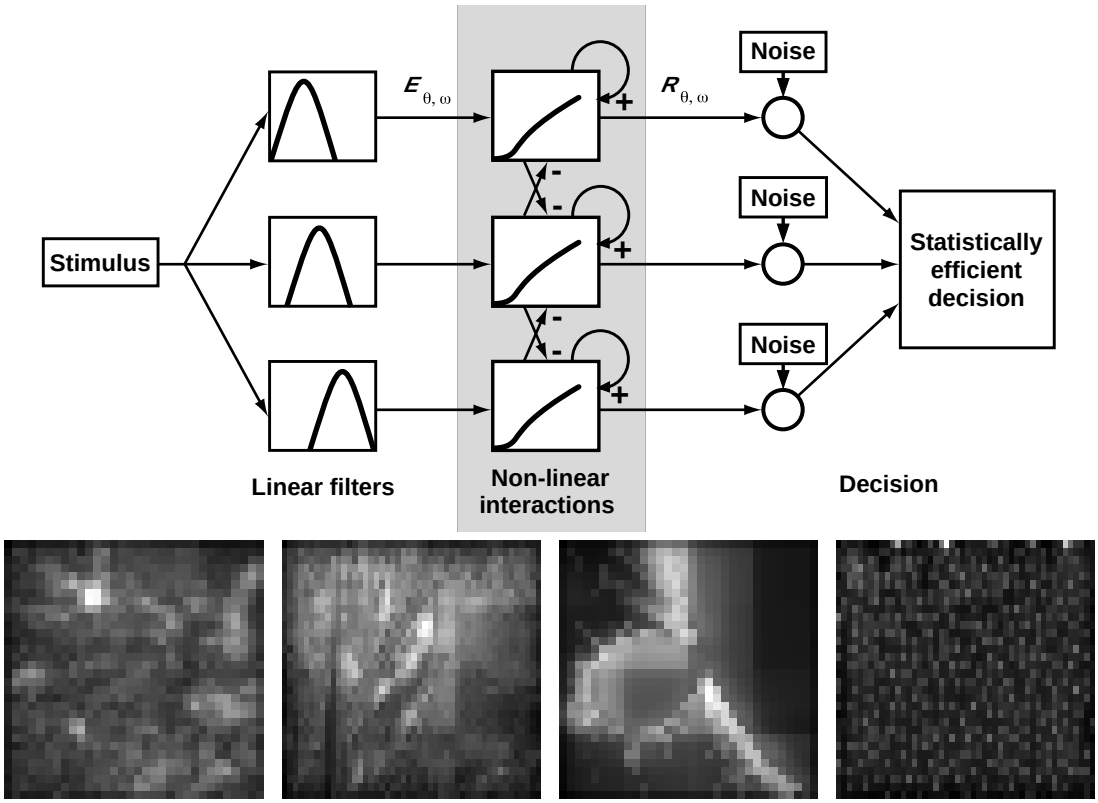


Figure 7.1. At top is a schematic diagram of the short-range orientation interactions model. In this model, an input image is passed through a set of linear filters tuned to different orientations and spatial frequencies. The linear outputs feed forward into a second stage, in which the set of filter outputs corresponding to a given spatial location form a pool that divisively inhibits each unit's response at that location. As a result of this recurrent processing, the second stage output exhibits gain control and contrast enhancement relative to the first stage. The bottom row of images shows the salience map produced by a modified model including short-range orientation interactions, for each of the four exemplar images from Figure 6.1. Figure adapted from Lee et al. [1999].

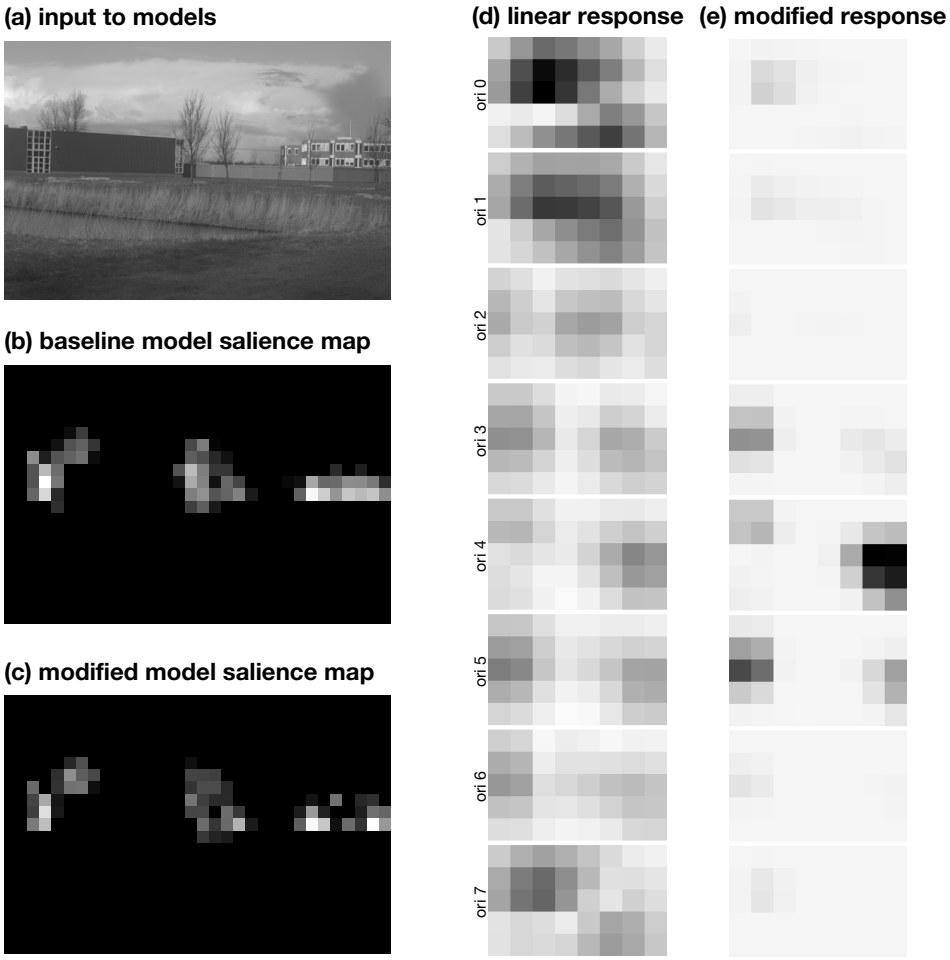


Figure 7.2. Depicted here is the effect of including short-range orientation interactions in the orientation channel of the salience model. The same input image **(a)** was given to the baseline salience model and the modified model including short-range orientation interactions, and the resulting salience maps from the two models are shown in **(b)** and **(c)**, respectively. At right **(d-e)** are shown the internal workings of the short-range orientation interactions for one of the nine spatial scales involved in the model. Each row represents one of eight filter orientations. In **(d)** is the raw linear response of the oriented filters—these responses are used directly in the baseline model. However, the modified model uses the results shown in the next column **(e)**, which include the effects of cross-scale and cross-orientation inhibition (even though only one scale is shown here, the results still reflect the effects of cross-scale inhibition from other scales that are not shown here).

7.3 Model fits

Table 7.1. Shown here is a comparison of the baseline salience model and the modified model including short-range orientation interactions. The models are judged by comparing their predictions with the scanpaths recorded during the free-viewing task (Figure 6.2), using the normalized scanpath salience (NSS) as described in the previous chapter. The first four rows are arranged identically to Table 6.1, except that a new has been added for the modified model introduced in this chapter. The bottom four rows express the performance of each model as a percentage of the NSS attained by the inter-observer model, as in Table 6.2.

	Outdoor	Fractal	Satellite	Gabor snake	Gabor array
NSS					
Random model	-0.01	-0.02	0.02	-0.01	0.02
Baseline salience model	0.69	0.44	0.62	0.10	0.14
Short-range interactions	0.75*	0.56*	0.71*	0.11	0.14
Inter-observer	1.30*	1.13*	1.10*	1.15*	0.91*
NSS % of Inter-observer NSS					
Random model	0%	-2%	2%	-1%	2%
Baseline salience model	53%	39%	57%	9%	15%
Short-range interactions	57%	50%	65%	10%	15%
Inter-observer	100%	100%	100%	100%	100%

$*$, models whose fi t was signifi cantly better than the corresponding baseline salience model, $p < 0.05$

When the model of short-range orientation interactions was substituted for the standard orientation channel in the salience model, we observed a statistically significant 10–20% improvement in the model fits across all of the image classes, except for the Gabor snake and Gabor array images in which there was no effect of the short-range orientation interactions (Table 7.1). Average NSS values ranged from 0.56 to 0.75 for the natural image classes, and from 0.11 to 0.14 for the Gabor arrays.

The interactions in the modified model were based on the lateral inhibition that takes place within a V1 hypercolumn, which in turn is an abstraction of the concept that for a given receptive field visual space, there is a confined population of cells in primary visual cortex that are tuned to all possible spatial scales and orientations. Lateral inhibition is a ubiquitous element in sensory processing along spatial, temporal, and higher-order feature dimensions, as it decorrelates the input, emphasizing regions with high contrast and deemphasizing relatively homogeneous regions. Ultimately this allows behaviorally relevant input to be repre-

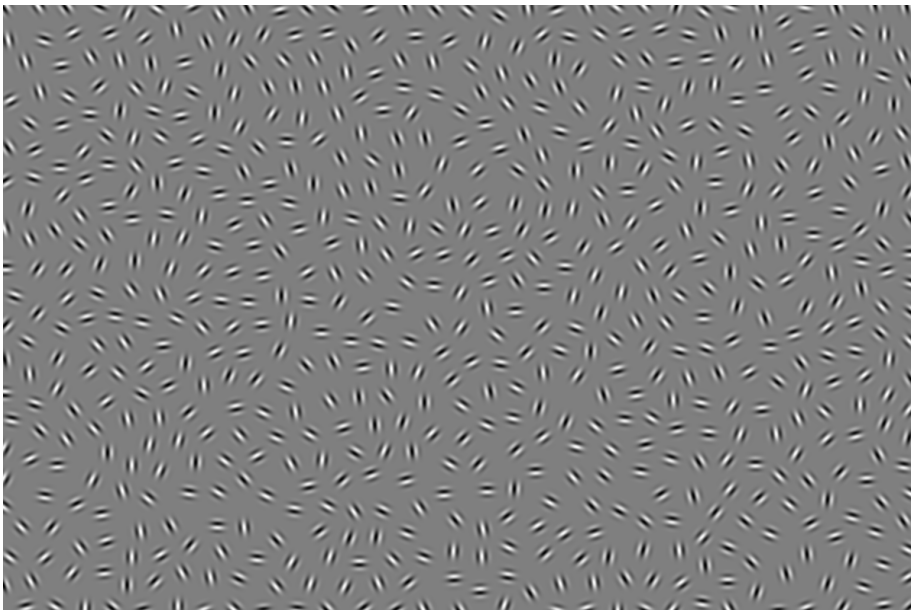
sented in a more explicit and compact manner. Lee et al. [1999] used psychophysics experiments to validate a V1 hypercolumn model, showing that changes in attentional state could be explained by changing the relative contributions of feed-forward excitatory and feed-back inhibitory connections. These connections determine, among other things, how easily an observer is able to identify a low-contrast grating in the presence of an overlapping grating of a different orientation. When we included these connections in our salience model, we found that the model's salience maps predicted observers' fixation locations significantly better. Thus, these connections, previously modeled with well-controlled minimalistic stimuli, also appear behaviorally relevant under less restrictive task conditions involving free-viewing natural scenes.

CHAPTER 8

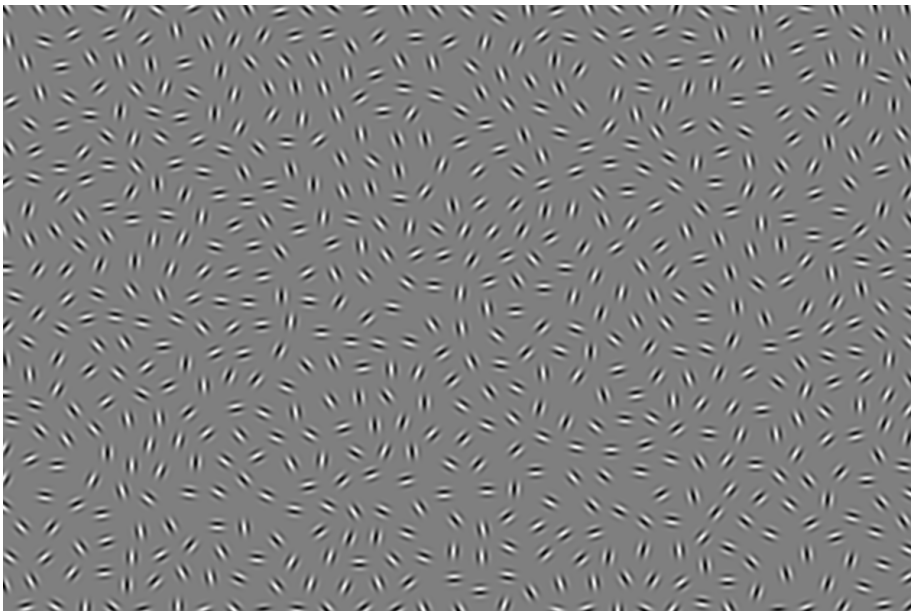
Long-range orientation interactions

8.1 Introduction

The visual system is exquisitely sensitive to contours, even when they are defined by scant evidence, as in the Gabor arrays shown in Figure 8.1. The ability to detect such contours is thought to rely on long-range interactions among V1 orientation-tuned units with non-overlapping receptive fields. The presence of these interactions has been inferred from neuroanatomy and electrophysiology [Blasdel, 1992, Pettet and Gilbert, 1992, Das and Gilbert, 1999, Stettler et al., 2002] and from psychophysical studies demonstrating increased or decreased contrast detection thresholds at a central location depending on the presence and orientation of surround elements [Polat and Sagi, 1993, 1994a, Zenger and Sagi, 1996, Zenger et al., 2000]. An appropriate arrangement of connection strengths [Polat and Sagi, 1994b, Li, 1998, Braun, 1999a, Li and Gilbert, 2002], involving facilitation between nearly collinear edge segments and inhibition between non-collinear parallel and orthogonal segments, has the effect of enhancing the activity of units that respond to the segments comprising an elongated contour. We adapted one model of such interactions [Mundhenk and Itti, 2002] to test whether such contour-facilitation plays a role in directing eye movements, and furthermore how that role depends on the relevance of contours to the behavioral task.



(a) Gabor array



(b) Gabor array with embedded implicit contour ("snake")

Figure 8.1. These are examples of the Gabor array images that were used in psychophysics experiments and in testing of the contour-integration model. The two arrays shown here are identical except for the orientation of certain elements in **(b)** giving rise to an implicit contour, or "snake." In each array, the spatial positions of the Gabor elements are assigned randomly subject to certain minimum and maximum inter-element distance constraints. The phases of the gratings are assigned randomly for all elements. The orientations are also assigned randomly except for the snake elements. In fact, the *only* feature that distinguishes a snake from its background is the orientation of its member elements; phase and inter-element spacing are identical for snake and background elements.

8.2 Contour-integration model

We adapted a model of long-range orientation interactions [Mundhenk and Itti, 2002] (Figure 8.2) that was included as a new channel in the salience model[1]. In essence, this model relies on a set of weight matrices that determine how one orientation-tuned unit is influenced by other such units at different distances and orientations (Figure 8.3), in a manner reflecting the long-range connections thought to be present in primary visual cortex [Blasdel, 1992]. These matrices are sometimes described by their shape which resembles a "butterfly" or "bow-tie," with wedges of excitatory connections leading from the central unit to other units that are similarly tuned and nearly collinear. Outside these wedge-shaped regions, there are inhibitory connections from the central unit leading to other similarly tuned units that are nearly parallel but not collinear. Our model did not include interactions among orthogonal or nearly orthogonal units.

Qualitatively, this model of contour integration was found to yield very satisfactory results when applied either to naturalistic images or to the Gabor arrays. Figure 8.4 shows the output from the contour model for two of the overhead satellite photos; the model highlights the contours that might be intuitively expected to be salient contours in such images, including roads, coastlines, and rivers. We also made a rough comparison between this contour model and two standard edge-detection algorithms from computer vision, the Canny and Sobel algorithms. These results are shown in Figure 8.5; it appears that the contour model is more able to highlight the "main contour" as our visual system perceives it, without allowing large amounts of background noise to seep through into the output. This is an admittedly weak test of the Canny and Sobel algorithms, since we used the implementations straight "out-of-the-box" from MATLAB; nevertheless it suggests that the contour model is perhaps qualitatively different from such algorithms.

[1]Special thanks to Nathan Mundhenk for sharing a reference implementation of the contour-integration model.

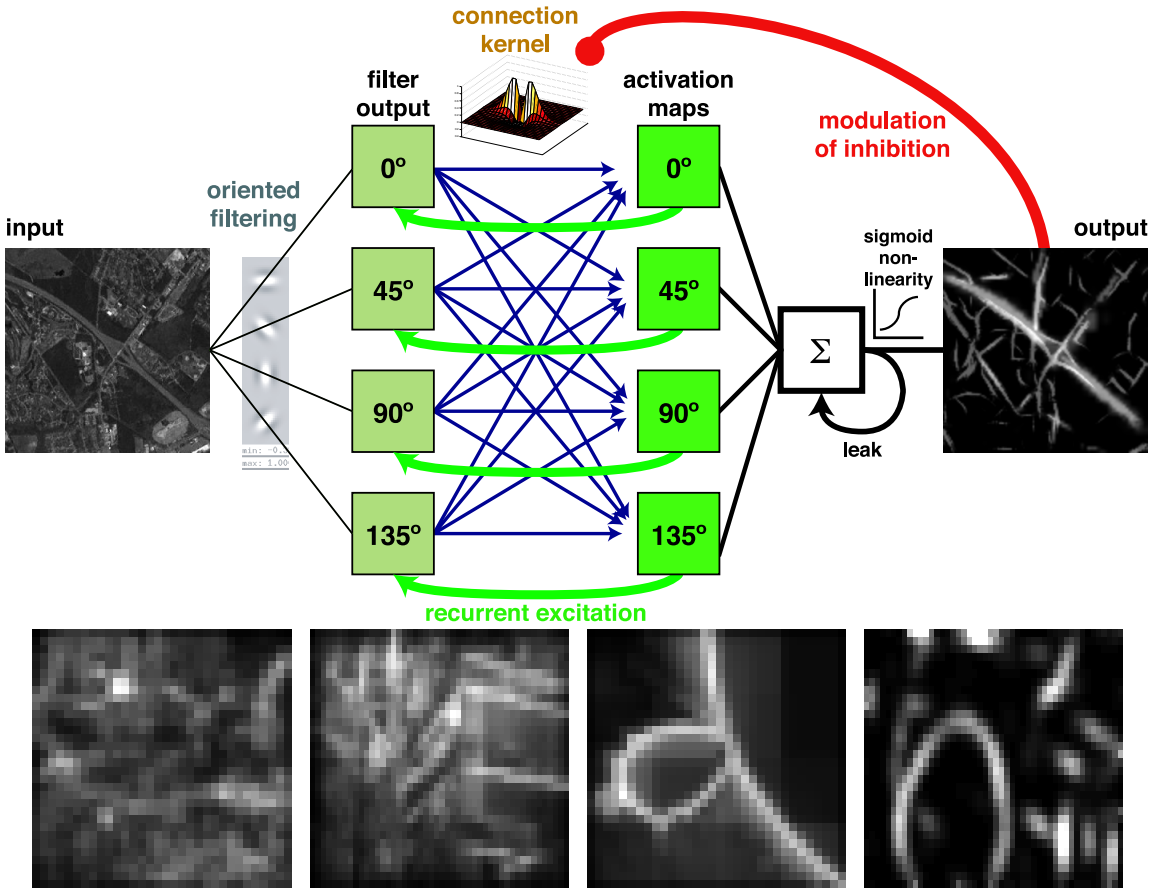


Figure 8.2. At top is a schematic diagram of the long-range orientation interactions (contour-integration) model. In this model, an input image is passed through a series of filters tuned to 12 orientations (only 4 are depicted in the figure), all tuned to the same spatial scale. The first stage filter outputs feed forward into second stage activation maps via a set of kernels that specify connection strengths as a function of relative spatial position and relative orientation tuning (see Figure 8.3). These connections are arranged so as to selectively enhance locations that form part of an elongated contour. The activation maps are summed across orientations and passed through a sigmoid non-linearity to yield the final output map. The model output evolves iteratively (three iterations were used in the present study); the second stage maps recurrently excite their first stage counterparts, and the output map recurrently dynamically modulates the strength of inhibition within the connection kernels in order to limit the dynamic range of the output. In practice, the model was instantiated for three spatial scales, but there were no interactions between scales at the intermediate stages; the outputs from each of the spatial scales were summed at the final stage to produce an overall output. The bottom row of images shows the salience maps produced by a modified salience model including a contour-integration channel, for the four exemplar images shown in Figure 6.1.

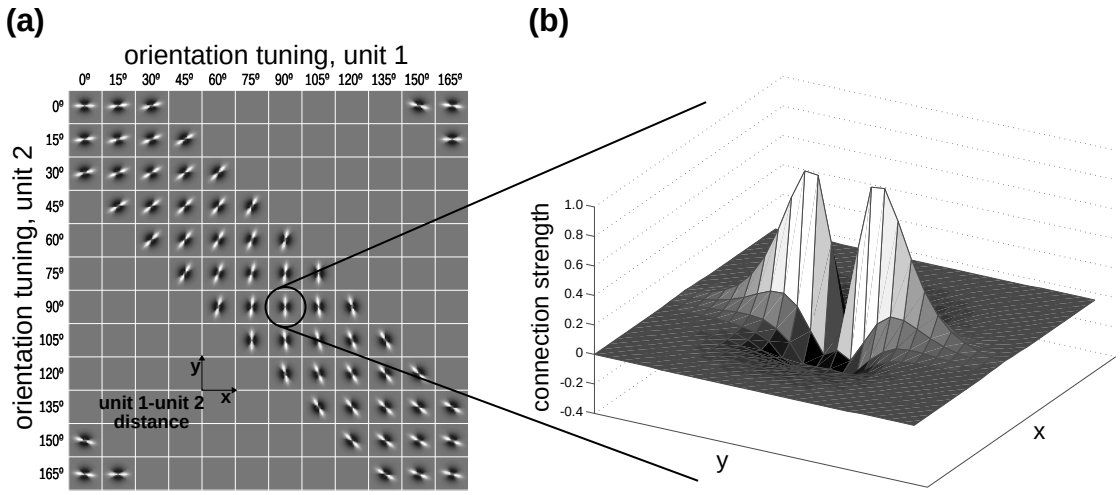


Figure 8.3. Illustration of the weight matrices that connect neighboring units at different orientations in the contour model of Figure 8.2. **(a)** Each grid entry is a spatial array depicting the connection strengths between a central unit (unit 1) tuned to the orientation given by the column label, and a neighboring unit (unit 2) tuned to the orientation given by the row label. Within each grid entry, the spatial separation between unit 1 and unit 2 is represented by the x- and y-axes indicate, and connection strength is represented by gray level: lighter pixels reflect regions of excitation, darker pixels reflect regions of inhibition, and gray pixels reflect the absence of any connection. The "butterfly" shape of the kernels indicates that there are symmetric cones of excitation connecting a central unit with neighbors whose position and orientation is such that the two units are either collinear or fall on a contour of low curvature, as well as symmetric flanks of inhibition between units that represent contour elements that are nearly parallel and non-collinear. **(b)** An enlargement of the $90^\circ/90^\circ$ kernel. Here, connection strength is represented by z-axis height as well as gray level, with values above and below 0.0 representing excitation and inhibition, respectively.

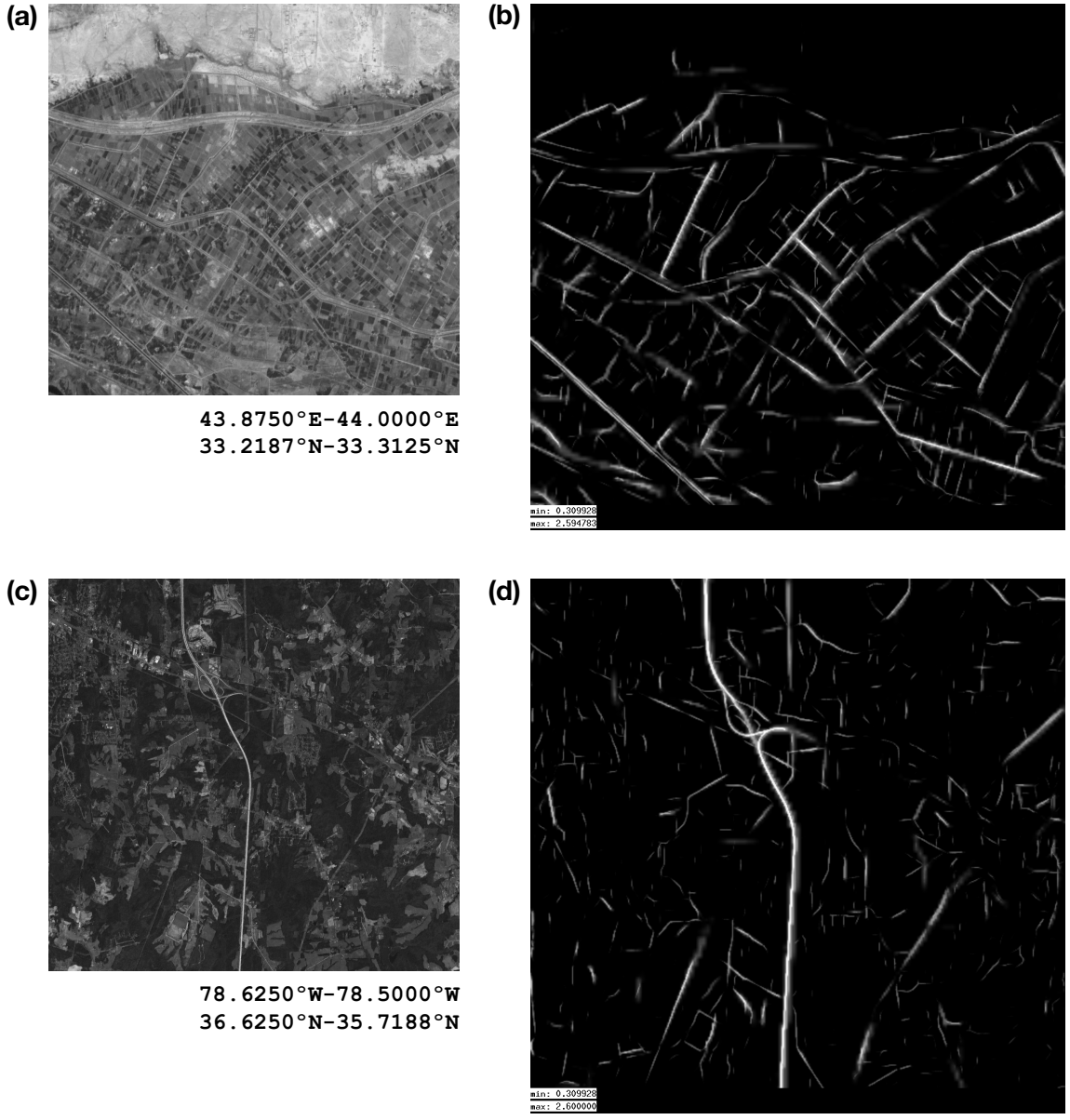


Figure 8.4. Shown here are two images **(a,c)** from the overhead imagery database, and the corresponding outputs from the contour-integration model **(b,d)**.

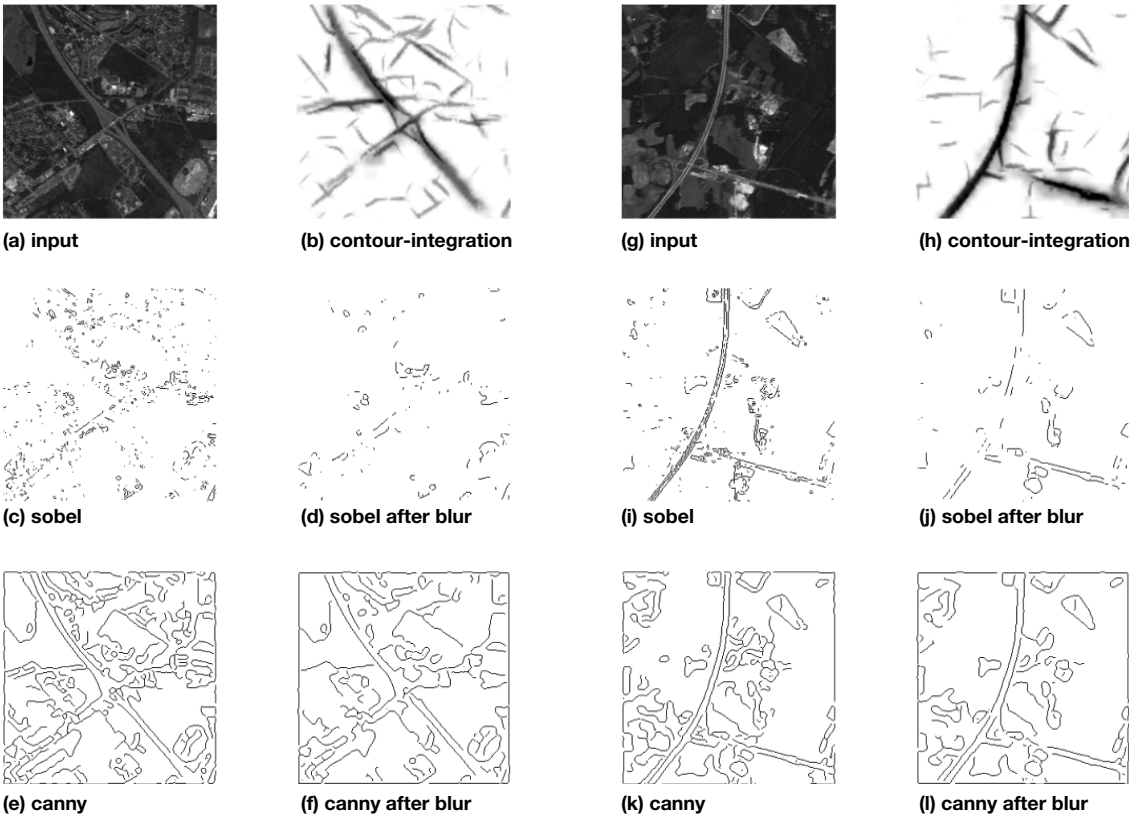


Figure 8.5. Shown here are two image fragments from the overhead imagery database, along with the results of processing those images with either the contour-integration model or standard edge-detection algorithms in the MATLAB image-processing toolbox (`canny` and `sobel`). **(a)** Input image. **(b)** Result of contour-integration model. **(c)** Result of Sobel edge-detection algorithm; **(d)** Sobel algorithm after first blurring the input image. **(e)** Result of Canny edge-detection algorithm; **(f)** Canny algorithm after first blurring the input image. **(g-l)** Same operations for a different input image.

8.3 Contour-detection task

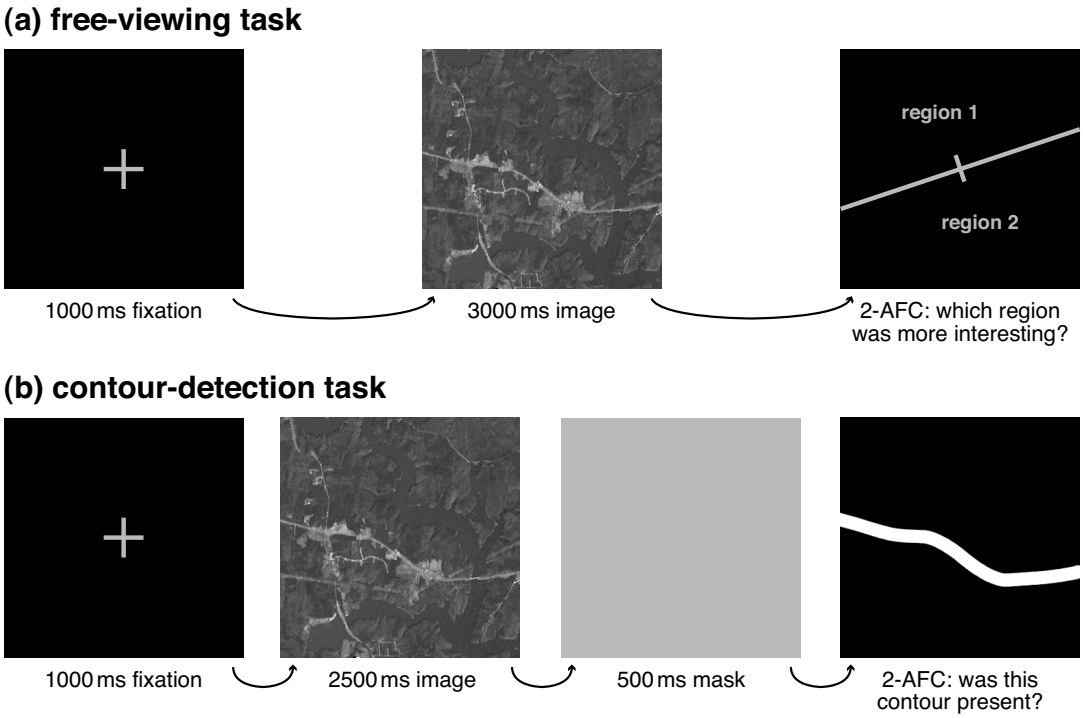


Figure 8.6. Illustration of the new contour-detection task **(b)**, in comparison with the original free-viewing task **(a)** (duplicated from Figure 6.2). In the contour-detection task **(b)**, each trial began with a fixation cross followed by a stimulus image as in the free-viewing task. However, when the image disappeared it was replaced by a full-screen uniform white mask. This was followed by a new response screen containing a single schematic contour, and subjects made a 2-AFC as to whether there had been a matching contour at the same location in the just-seen image (in the example shown here, the contour does match the image). On 50% of trials there was such a match, and on the other 50% of trials a non-matching contour was selected from among the contours that matched other images in the same category.

We used a second task to investigate the influence of contours on fixation locations (Figure 8.6b). The overall format was similar to the free-viewing task, except (1) the image presentation time was shortened from 3000 ms to 2500 ms, (2) a full-screen uniform white mask was presented for 500 ms immediately after each image to prevent subjects from relying on retinal afterimages to perform the task, and (3) a different response was required, as explained next. After the image and the mask, subjects were presented with a schematic line-drawn contour, and were asked to respond with a key press to indicate whether that contour matched a contour that was present at the same location in the image they had just seen. On half of the trials, the contour was in fact a match to the preceding image, and on the other half,

the contour was a non-match (selected from a pool of contours that matched other images in the experiment). The schematic contours were Bezier curves that closely approximated the shapes of hand-picked salient contours in the target images.

In order to make within-subject comparisons between fixation locations for the same image but in different tasks (free-viewing or contour-detection), we used the following experimental design. Images were split into two batches, A and B. Subjects were split into two groups. Subjects in group one performed four blocks: free-viewing (A images), contour-detection (B images), contour-detection (A images), free-viewing (B images). Thus, these subjects were naive to the contour-detection task when performing the free-viewing on batch A, but not when performing the free-viewing on batch B. The primary comparison of interest is between the data collected in blocks 1 and 3 — that is, free-viewing versus contour-detection for the same subjects and the same images, with subjects' free-viewing behavior not tainted by knowledge of the contour task. A secondary comparison between blocks 4 and 2 allows us to address the effect of exposure to the contour task on subjects' behavior in the free-viewing task. Subjects in group two performed the tasks in the same order, but were shown image batch A in place of B and vice versa.

8.4 Model results

We added to the salience model a new channel for contour integration via long-range orientation interactions[2]. This led to improved model fits over the baseline salience model by 19–36% for the image classes, and by $\approx$ 300% for the Gabor arrays (see Table 8.1). Notably, only with long-range orientation interactions did the model's performance rise above chance levels for the Gabor arrays. In addition, for all image classes except the outdoor photos, contour integration led to a significant improvement in the fits beyond those attained by including the short-range orientation interactions. Turning again the theoretical upper limit on model performance attained by the NSS attained by the inter-observer model (defined in Section 6.6), we find that the modified salience model including a contour-integration channel reaches 36% to 74% of this maximum across the different image classes.

Further investigation revealed a large variability in the optimal strength of the

[2]The new channel was added in parallel to the existing luminance, orientation, and color channels. Another possibility we considered was using the contour channel as a multiplicative gate on the standard orientation channel, or even on the entire output of the salience model. In practice, we found that this arrangement offered no better fit with human fixation locations than did the standard arrangement of placing the contour channel in parallel next to the other channels

Table 8.1. This table extends the results shown previously in Tables 6.1, 6.2, and 7.1, now including the normalized scanpath salience (NSS) values resulting from the comparison of eye-tracking scanpaths with the modified model including a contour-integration channel.

	Outdoor	Fractal	Satellite	Gabor snake	Gabor array
NSS					
Random model	-0.01	-0.02	0.02	-0.01	0.02
Baseline salience model	0.69	0.44	0.62	0.10	0.14
Short-range interactions	0.75*	0.56*	0.71*	0.11	0.14
Contour-facilitation	0.72	0.60*	0.81*	0.41*	0.52*
Inter-observer	1.30*	1.13*	1.10*	1.15*	0.91*
NSS % of Inter-observer NSS					
Random model	0%	-2%	2%	-1%	2%
Baseline salience model	53%	39%	57%	9%	15%
Short-range interactions	57%	50%	65%	10%	15%
Contour-facilitation	55%	53%	74%	36%	58%
Inter-observer	100%	100%	100%	100%	100%

$*$, models whose fi t was signifi cantly better than the corresponding baseline salience model, $p < 0.05$

contour channel relative to the other channels, depending on the image class. The outdoor images were fit best with a relatively weak contour channel, the fractal images with an intermediate strength, and the satellite imagery with a relatively strong contour channel.

We used a second task to specifically address the role of elongated contours in selecting fixation locations, by asking subjects to view the same images under two different task conditions: first, the standard free-viewing task, and subsequently, a contour-detection task. Table 8.2 shows the results of comparing models with behavior in these two tasks. Overall, model performance was worse in predicting fixation locations in the contour-detection task than in the free-viewing task; this is likely because performing the contour-detection task involves a greater top-down component, whereas the model mimics only bottom-up components. Nevertheless, there was an interaction between task and model: the relative improvement due to the contour-integration model over the baseline model was greater for the contour-detection task than for the free-viewing task. That is, the contour-integration model was better suited to the contour-detection task.

Table 8.2. Shown here is a summary of the fits between each model and the eye-tracking data from the free-viewing and contour-detection tasks (Figure 8.6), using the normalized scanpath salience metric described in Figure 6.7 and Table 6.1. In the free-viewing task, subjects passively observed images, while in the contour-detection task, subjects were presented with a schematic contour following each image and were required to indicate whether that contour matched one that was present in the just-seen image. These data illustrate that, relative to the baseline model, the modified model including contour-integration was more predictive of fixations in the contour-detection task (row 6 versus row 2) than of fixations in the free-viewing task (row 5 versus row 1).

Model	Task	Gabor array	Gabor snake	Outdoor	Satellite
Baseline salience model	free-viewing	0.136	0.145	0.450	0.398
	contour-detection	0.117	0.126	0.505	0.216
Short-range interactions	free-viewing	0.122	0.152	0.538	0.559
	contour-detection	0.135	0.131	0.585	0.364
Contour-integration	free-viewing	0.565	0.521	0.512	0.713
	contour-detection	0.619	0.540	0.588	0.481

8.5 Discussion

We have described in this chapter a model a second type of connection in primary visual cortex, that of orientation-dependent long-range connections between different hypercolumns. Such connections or their computational equivalent have been introduced to explain the subjective salience of implicit contours like Gabor "snakes" that would otherwise be invisible to purely local processing. Indeed, without long-range connections, the salience model performed very poorly in predicting observers' fixation locations in the Gabor arrays, since each individual Gabor element appears equally salient to a purely local mechanism. As we expected, the model performance increased dramatically (more than threefold) when the long-range connections were included. However, somewhat unexpected was the fact that these connections lead to more modest improvements in predicting fixation locations in the natural image categories. This could be explained in one of two ways: either the model was not accurately identifying what observers' considered to be "contours," or the observers were giving relatively little weight to the contours that were present. To distinguish between these possibilities, we performed a second psychophysics experiment in which observers viewed images under two different task conditions, one requiring them to specifically attend to contours, and one requiring only free viewing. If our model of contour integra-

tion based on long-range connections was simply inaccurate, then it should not have shown any additional benefit in predicting observers' contour-detection behavior over their free-viewing behavior. Instead, we found that the improvement in model fit due to contour-integration was greater when subjects performed the contour-detection task than when they performed the free-viewing task. This implies that, although our contour-integration model was accurately highlighting what would qualitatively be identified as "contours," observers' fixation locations were only weakly influenced by the presence of elongated contours, at least in natural images where other salient image features were present.

CHAPTER 9

Mouse-clicking

9.1 Introduction

The well-established link between overt eye movements and covert attentional shifts makes eye tracking an appropriate tool for investigating visual attention; yet, the logistical requirements of eye tracking—requiring subjects to participate in a particular physical location, with an experimenter present at all times—place practical limits the amount of data that can be acquired. With alternative approaches, we can make a different trade-off, gaining access to a much larger subject pool and data source, in return for allowing for somewhat less well-controlled experimental conditions and a less direct connection between the behavioral modality and the underlying visual attention mechanisms. In concrete terms, our free-viewing eye movement task could be replaced with a task in which subjects indicate "which locations are interesting" by, for example, pointing a finger or making a verbal report [Astafiev et al., 2003]. To validate this type of method, we used one such approach involving "mouse-clicking." [1] Our results indicate clear similarities as well as intriguing differences in the pattern of locations resulting from mouse-clicking and eye-tracking sessions. Following the success of this pilot study, a wide variety of full-fledged experiments are now possible that would be far more difficult in a conventional eye-tracking setting.

[1] The work described here was performed in collaboration with Dr. Chris Scheier of MediaAnalyzer, Inc. (`http://www.mediaanalyzer.com`), where the internet application was developed and where the data-tracking database was hosted

9.2 Mouse-clicking: an alternative to eye tracking

Subjects viewed images on a computer screen in the same task setup as in the eye-tracking experiments. Instead of having their eye movements recorded, subjects used a standard computer mouse to click on points in the image that seemed to attract their attention. To further capitalize on potential data sources, this task was implemented as an application that could be run on a internet web browser, so that subjects could perform the task at a time and place of their own choosing. Once subjects completed the task, their results were sent returned over the network to a central database server for storage and later analysis.

When subjects logged on to the website for the experiment, they were taken through a brief (5-minute) training sequence to familiarize them with using the mouse to click in images, and to introduce some consistency in the manner in which different subjects clicked in the images. The goal was for subjects to think of the mouse as simply an extension of their eyes, so that they would simply look around naturally in the image, and "let the mouse follow their eyes." This is somewhat in contrast to the way that experienced computer users are accustomed to using the mouse, in which each mouse click is the result of a deliberate action with a specific desired result; therefore, a bit of unlearning was required for such subjects to be able to treat each mouse click with less forethought. These instructions had been honed previously at MediaAnalyzer for their own uses in analyzing how people view advertisements, websites, and other marketing devices. At the beginning of the training sequence, subjects were taught to simply press the mouse button repeatedly and rapidly for periods of several seconds at a time, maintaining a rate of at least 3–4 clicks per second. Next, subjects were taught to scan an image and, within a limited time, click on targets that met certain predefined criteria (these pages showed drawings of different food items with different price tags attached; subjects were asked to click on the expensive items costing more than $100 each). Finally subjects were shown several sample images in a format similar to that used for the main experimental sequence, and asked to view the image naturally and click at locations that attracted their eye movements.

The main experimental sequence was the same as in the eye-tracking task: each trial consisted of a central fixation cross followed by the stimulus image. Unlike the eye-tracking task, we had no control over the size of the computer screen used to view the images nor over the distance from which the subject observed the screen, since subjects performed the task on their personal computers. We used 10 images

Table 9.1. Shown here is a comparison between the fits of model predictions to fixations from eye-tracking tasks, and clicks from the mouse-click task, for each of the model variants introduced in the previous three chapters. The first five rows give the average normalized scanpath salience (NSS), quantifying the degree of correspondence between a series of spatial locations (eye fixations or mouse clicks) and the salience map predicted by one of the models. The next five rows give each model's performance as a percentage of the theoretical maximum, defined by the NSS achieved by the inter-observer model.

	Outdoor		**Fractal**		**Satellite**	
	fixations	*clicks*	*fixations*	*clicks*	*fixations*	*clicks*
NSS						
Random model	-0.01	0.00	-0.02	0.00	0.02	0.03
Baseline salience model	0.69	0.43	0.44	0.84	0.62	0.73
Short-range interactions	0.75*	0.40	0.56*	0.85	0.71*	0.72
Contour-facilitation	0.72	0.40	0.60*	0.93*	0.81*	0.77*
Inter-observer	1.30*	1.27*	1.13*	1.54*	1.10*	1.07*
NSS % of Inter-observer NSS						
Random model	0.00%	0.00%	-0.02%	0.00%	0.02%	0.03%
Baseline salience model	0.53%	0.34%	0.39%	0.54%	0.57%	0.68%
Short-range interactions	0.57%	0.31%	0.50%	0.55%	0.65%	0.67%
Contour-facilitation	0.55%	0.31%	0.53%	0.61%	0.74%	0.72%
Inter-observer	1.00%	1.00%	1.00%	1.00%	1.00%	1.00%

$*$, models whose fi t was signifi cantly better than the corresponding baseline salience model, $p < 0.05$

from each of the overhead imagery, outdoor photos, and fractal images databases; 81 subjects completed the overhead imagery section, 43 completed the fractals, and 100 completed the outdoor photos.

9.3 Salience model fits with mouse clicks

Since we have data from many more subjects in the mouse-click task than in the eye-tracking task, it is easier to judge the qualitative patterns of responses in the mouse-click data. Figures 9.1, 9.2 and 9.3 show several example images from each image database, along with pooled responses from all subjects in the two tasks. In the mouse-click data it is easier to visually identify significant clusters of responses; for example Figure 9.1a reveals several clusters around apparent roads and inhabited regions.

We compared the mouse click sequences with the salience maps predicted by the different variations of the salience model introduced over the last three chap-

ters, using the normalized scanpath salience (NSS); these results are shown in Table 9.1. In addition, we compared the NSS values observed for the mouse clicks with those observed for the eye fixations in the eye-tracking tasks. As with the eye-tracking results, we find that the baseline salience model gives NSS values for the mouse clicks that are far above chance levels, ranging from 0.43 to 0.84. On the other hand, the short-range orientation interactions, which produced significantly higher NSS scores for each of the three image classes in the eye-tracking experiment, now produces no significant effect for any of the image classes in the mouse clicking experiment. The contour-integration model continues to lead to a significantly increased NSS score in the mouse click results for fractal and overhead satellite images, the same two categories for which it produced a significant increase in the NSS score in the eye-tracking results. Overall, the mouse clicks are less well-predicted by the models in the outdoor images, but are somewhat better-predicted by the models in the fractal and satellite images.

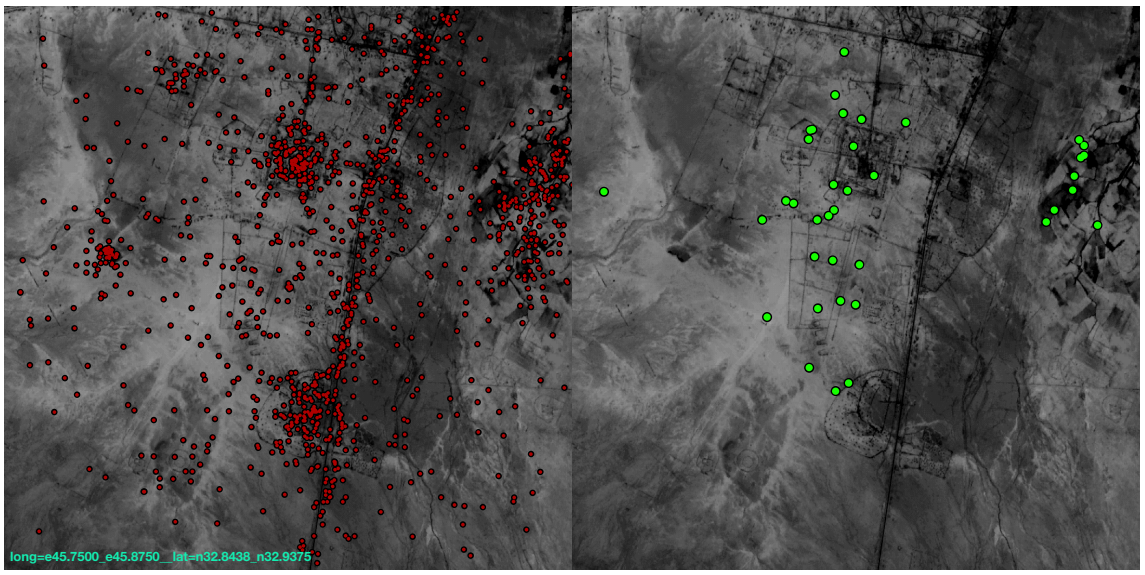
The relationships between the eye-tracking and mouse-clicking results suggest that similar, but clearly distinct, mechanisms underlie each of the two behavioral modalities. Our premise has been that a common attention-shifting mechanism, perhaps driven by a salience map, is responsible for both behaviors. Yet differing top-down attentional states are known to have dramatic effects on visual processing; Lee et al. [1999] showed that the differences in psychophysical performance between high-attention and low-attention conditions in the periphery could be explained by significant dynamic changes in the strength of self-excitation and inhibition. The model of short-range orientation interactions described in Chapter 7 was based on the high-attention parameters from Lee et al. [1999], and this channel was shown to have a positive influence on the salience model's ability to predict eye movements. In contrast, we have seen here that it has no such influence on the model's ability to predict mouse clicks. Thus something has changed in the processing sequence from visual to motor output that has changed which regions are selected for closer inspection (either by an eye movement or by a mouse click). Two possibilities arise:

- the internal salience map is different in the two cases, perhaps due to top-down influences based on the different tasks contexts of the eye-tracking and mouse-clicking experiments; or

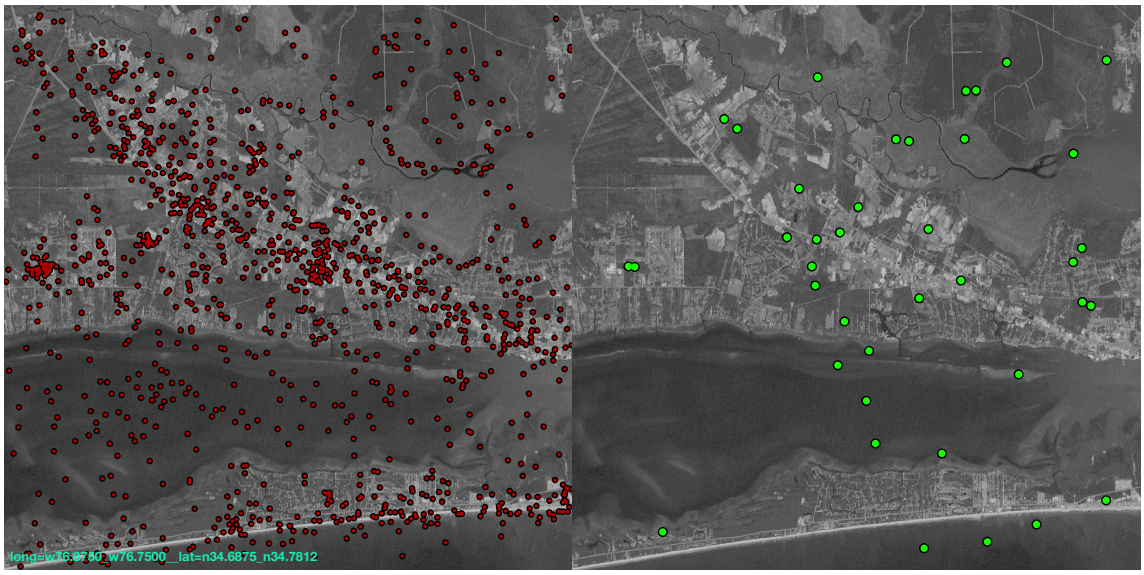
- the internal salience map is computed identically in both cases, and the two behaviors rely involve different mechanisms for selecting the next target.

Or very likely, a combination a both. The first possibility is supported by the differences in psychophysical performance due to attentional state [Lee et al., 1999], and additionally by the fact that the mouse-click experiments were performed under less well-controlled conditions than the eye-tracking experiments (*e.g.*, subjects were not isolated from distractions, their computer screens may not have been placed in optimal lighting or at an optimal distance). The second possibility is supported by the fact that eye movements and mouse clicking are likely to have very different cost metrics for selecting a new fixation or click target, given the location of the current target; it is very "cheap" to make an eye movement from one side of screen to the other, while it is relatively "expensive" to move the mouse the same distance (particularly if one is using the type of trackpad or pointing stick common on current laptop computers). Thus an optimal eye movement scanpath might involve large skips around the image, putting the priority on visiting all of the most salient regions as soon as possible; on the other hand, an optimal mouse-click sequence might be more measured, temporarily ignoring potentially salient but distant targets in favor of fully exploring the current target area before moving on to a new location.

In a variant approach using mouse clicks used by Parkhurst and Niebur [2003], subjects first freely view each image as in the eye-tracking experiments, and then *after* the trial was complete, click on "the five most interesting locations," for example. This avoids the possibility of subjects' free-viewing of the images being contaminated by simultaneous performance of the mouse clicking, which might, for example, distract subjects from the scanpath that they might otherwise make. On the other hand, waiting until the end of the trial to collect the mouse click responses has a strong disadvantage in the context of our goal of studying bottom-up attention, in that a much larger top-down component is likely to enter into subjects' decisions at the end of the trial. For example, after having viewed the image for several seconds, subjects are likely to have formed certain declaration conclusions about the image ("there is an airport at the lower left," or "a woman is running down the path") and their decisions about which locations are interesting are likely to be influenced by their declarative knowledge ("are there any planes approaching the airport?" or "where is the woman going?"). In contrast, by collecting mouse clicks (or eye movements) while subjects are in the process of viewing the image, we instead gather information about which locations first attracted their attention (and perhaps ultimately led them to their declarative conclusions about the image).

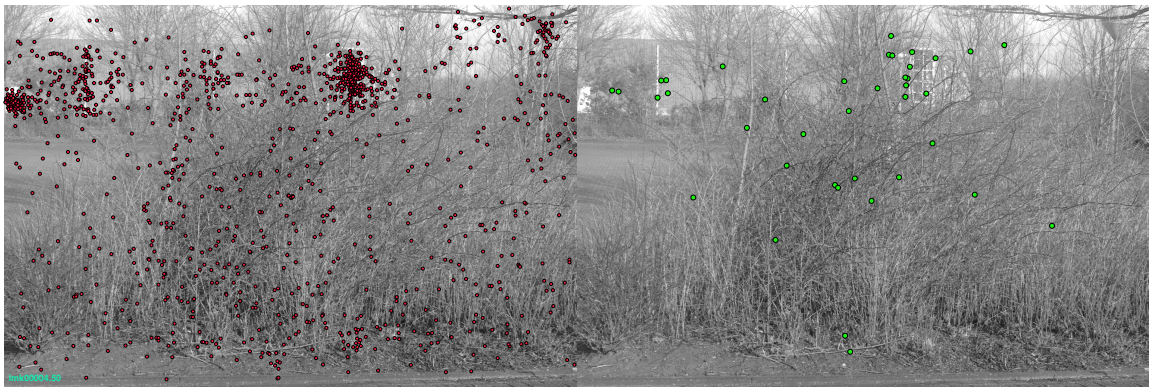


(a) mouse clicks **(b) eye fixations**

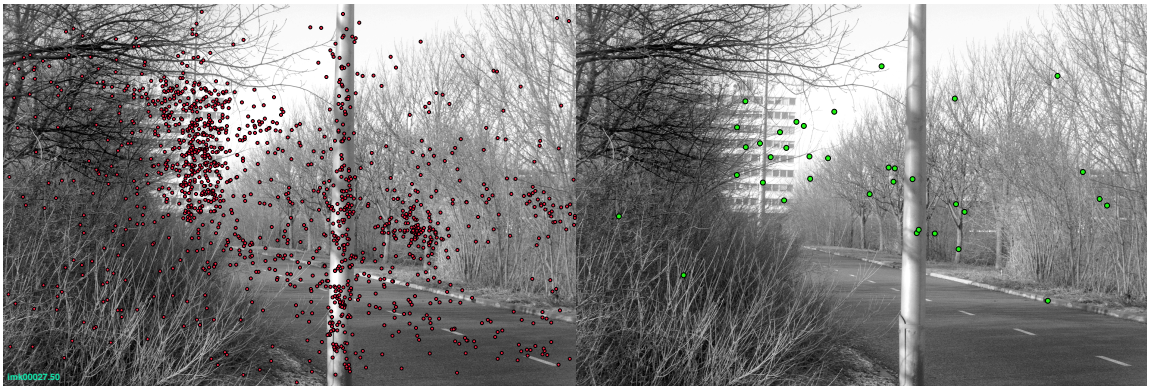


(c) mouse clicks **(d) eye fixations**

Figure 9.1. Mouse clicks **(a, c)** from 81 subjects, and eye fixation locations **(b, d)** from 4 subjects, for two sample images from the overhead satellite imagery database.

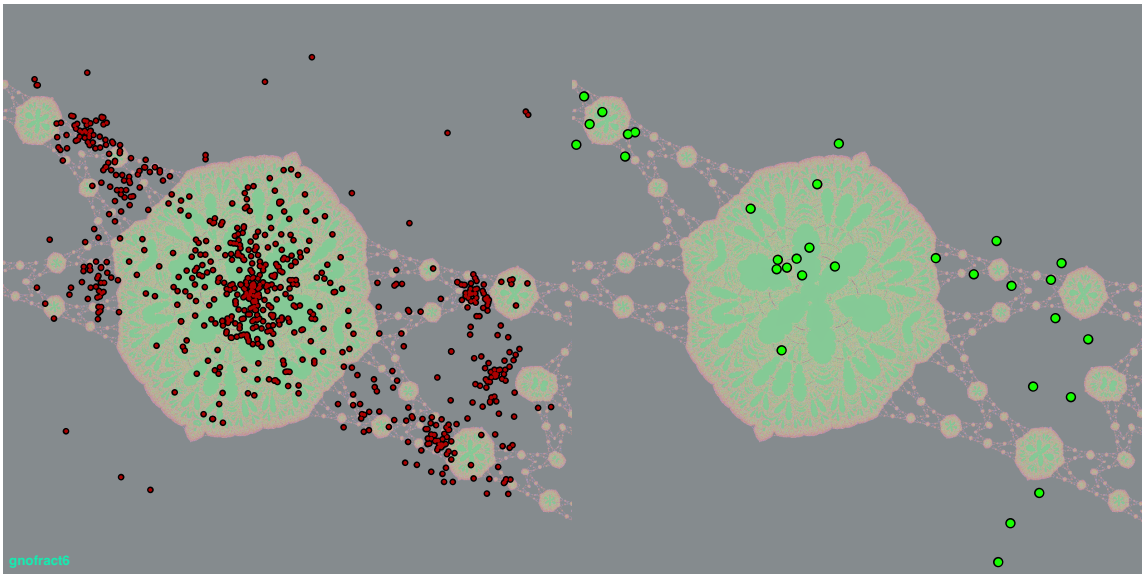


(a) mouse clicks **(b) eye fixations**

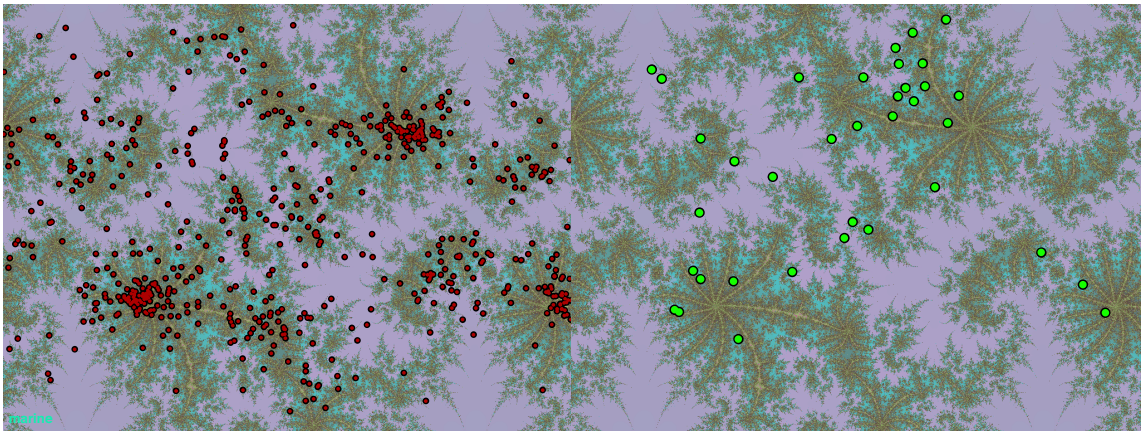


(c) mouse clicks **(d) eye fixations**

Figure 9.2. Mouse clicks **(a, c)** from 100 subjects, and eye fixation locations **(b, d)** from 4 subjects, for two sample images from the outdoor photos database.



(a) mouse clicks **(b) eye fixations**



(c) mouse clicks **(d) eye fixations**

Figure 9.3. Mouse clicks **(a, c)** from 43 subjects, and eye fixation locations **(b, d)** from 4 subjects, for two sample images from the fractals database.

CHAPTER 10

Conclusion

The working goal throughout this thesis has been to develop computational models that are not required to excel in any real-world task, except to faithfully mimic the behavior of human vision. Of course, human vision works quite well most of the time, so computational models that meet our stated goal may very well end up being well-suited to a number of real-world tasks. This is the crossroads of neuroscience-enabled machine vision, where biology meets engineering.

What have we learned about biology?

- In Chapter 2 we introduced three sets of schematic stimuli that are well suited for computational studies of the categorization process. We showed that at least one of these types of stimuli yields neural activation in the same brain area (the fusiform face area) that known to be specifically activated by images of real faces.

- In Chapter 3 we explored a number of existing computational models of categorization, as well as a new roaming exemplar model, and found that, contrary to the potential implications of existing models, human categorization behavior need not rely on an unlimited memory capacity; on the contrary, for the simple stimuli we used in these experiments, the models that most parsimoniously explained subjects' performance had fewer, rather than more, stored memory traces.

- In Chapter 4 we saw, using multidimensional scaling (MDS), that human observers are remarkably capable of acquiring internal representations for

parametric stimuli that are highly similar to the representations that we, as experimenters, used to design the stimuli (but which were unknown to the subjects). We also found that two different tasks for building such representations produce similar results, and this will allow for future comparisons between studies in which human subjects have performed one task, and monkey subjects have performed the other.

- In Chapter 5 we used a simple model of early vision to demonstrate a potential mechanism by which the visual system might acquire the kind of intermediate representation observed with MDS. This model was not specifically tuned to perform categorization tasks, but the intermediate representations generated by this model could be used to drive the categorization models and match human behavior to within 10% of the accuracy achieved by the original categorization models.

- In Chapter 6 we introduced several classes of images for use in eye-tracking (and later, mouse-clicking) experiments, and introduced a salience model framework for predicting fixation locations. We showed that the baseline salience model is able to predict fixation locations with far better than chance accuracy, in agreement with previous results. In absolute terms, we can say that the salience model can predict a subject's fixation locations over a period of three seconds with roughly half the accuracy as would be achieved if the predictions were instead based on *other* subjects' fixation locations.

- In Chapter 7 we included into the salience model a set of short-range inhibitory interactions among orientation-tuned units, and showed that this modified model was significantly better at predicting fixation locations than was the baseline model. Thus the neural mechanisms implied by the short-range interactions, previously supported by evidence from psychophysical performance with highly-controlled but artificial stimuli, are now also supported by psychophysical performance in a less-restrictive task (free viewing) involving naturalistic stimuli.

- In Chapter 8 we tested a model of contour integration based on long-range interactions among orientation-tuned units, included as a new channel in the salience model. We found that this new channel was most useful in predicting the locations fixated by subjects in a task requiring attention to contours, and was less useful in predicting the locations fixated in a simple free-

viewing task. Thus, although the model was accurately identifying contours, we found that these contours are only weakly salient for human observers unless there are specific reasons to attend to contours.

- Finally, in Chapter 9 we introduced mouse-clicking as an alternative to eye tracking. This method allows people to participate in psychophysics experiments from the privacy of their own web browsers, allowing us access to a much larger subject pool. We compared the abilities of the different salience model versions to predict fixation locations and mouse-click locations. While in general the models were far above chance accuracy in predicting both types of data, there were some notable differences. In particular, short-range orientation interactions seem to offer no advantage in predicting mouse click locations, in contrast to the significant improvement they provide in predicting fixation locations. This suggests differences between the two tasks in the neural mechanisms involving the salience map, either in the construction of the map, or in the readout from the map, or a combination of both.

In addition to building our understanding of biological vision, we have strived to develop computational algorithms that are efficient enough to be useful in real-world machine vision applications. For example, the models of short-range orientation interactions and eccentricity-dependent filtering described here have efficient implementations that have only minimal impact on the execution time of the salience model, yet lead to significant improvements in the model's ability to match human behavior. In contrast, the model of contour-integration requires roughly an order of magnitude more processing time and is weakly relevant to behavior in some task conditions, but is also critically important in predicting behavior under other conditions such as the Gabor snakes that we tested, and also potentially in real-world tasks like road-finding in overhead imagery. Taken together, this suggests that a machine vision implementation might best compute an initial salience map based on local features alone, and secondarily perform more computationally intensive tasks like contour-integration or object-recognition within a restricted window selected by the first stage. Such systems will ultimately be useful both as stand-alone applications and as semi-automated assistants in tasks that rely on a human executor. The interface between biology and engineering is rich in research directions that will lead us closer not only toward understanding the inner workings of vision, but also toward building machines that assist, interact, collaborate, and synergize with real human visual systems.

BIBLIOGRAPHY

J.R. Anderson. The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429, 1991. 3.7.4

P. Arabie, S.M. Kosslyn, and K.E. Nelson. A multidimensional scaling study of visual memory of 5-year olds and adults. *Journal of Experimental Child Psychology*, 19:327–345, 1975. 4.2.2, 4.5

F.G. Ashby, editor. *Multidimensional Models of Perception and Cognition*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1992a. 2.1

F.G. Ashby. *Multidimensional Models of Perception and Cognition*, chapter 16: Multidimensional models of categorization. Lawrence Erlbaum Associates, Hillsdale, NJ, 1992b. 3.3, 4.1

F.G. Ashby and L.A. Alfonso-Reese. Categorization as probability density-estimation. *Journal of Mathematical Psychology*, 39(2):216–233, 1995. 3.7

F.G. Ashby and S.W. Ell. The neurobiology of human category learning. *Trends in Cognitive Sciences*, 5(5):204–210, 2001. 2.1, 3.3.2, 3.7.4, 3.7.5

F.G. Ashby and R.E. Gott. Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14:33–53, 1988. 3.3.3

F.G. Ashby and W.T. Maddox. Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, 37: 372–400, 1993. 2.1, 3.3.1, 3.3.3, 3.7, 3.7.2, 4.1

F.G. Ashby and E.M. Waldron. On the nature of implicit categorization. *Psychonomic Bulletin & Review*, 6(3):363–378, 1999. 2.1, 3.1, 3.3.2, 3.7, 3.7.4

F.G. Ashby, E.M. Waldron, W.W. Lee, and A. Berkman. Suboptimality in human categorization and identification. *Journal of Experimental Psychology-General*, 130 (1):77–96, 2001. 3.7

S.V. Astafiev, G.L. Shulman, C.M. Stanley, A.Z. Snyder, D.C. Van Essen, and M. Corbetta. Functional organization of human intraparietal and frontal cortex for attending, looking, and pointing. *Journal of Neuroscience*, 23(11):4689–4699, 2003. 6.7, 9.1

M.S. Beauchamp, L. Petit, T.M. Ellmore, J. Ingeholm, and J.V. Haxby. A parametric fMRI study of overt and covert shifts of visuospatial attention. *Neuroimage*, 14 (2):310–321, 2001. 6.1

G.G. Blasdel. Orientation selectivity, preference, and continuity in monkey striate cortex. *Journal of Neuroscience*, 12(8):3139–3161, 1992. 8.1, 8.2

I. Borg and P. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer-Verlag, New York, 1997. 4.2.1, 4.2.2, 4.3

J. Braun. Contour salience and striate cortex: A new model matches human sensitivity. *Investigative Ophthalmology & Visual Science*, 40(4):S780–S780, 1999a. 6.2, 8.1

J. Braun. On the detection of salient contours. *Spatial Vision*, 12(2):211–225, 1999b. 6.2

K.A. Briand, A.L. Larrison, and A.B. Sereno. Inhibition of return in manual and saccadic response systems. *Perception & Psychophysics*, 62(8):1512–1524, 2000. 6.7

E. Brunswik and L. Reiter. Eindruckscharaktere schematisierter gesichter (impression characteristics of schematized faces). *Zeitschrift fuer Psychologie*, 142:67–134, 1937. 2.2.1

D. Collett. *Modelling Binary Data*. Chapman & Hall/CRC, Boca Raton, 1991. 3.4

A. Das and C.D. Gilbert. Topography of contextual modulations mediated by short-range interactions in primary visual cortex. *Nature*, 399(6737):655–661, 1999. 8.1

S. Edelman. *Representation and Recognition in Vision*. MIT Press, Cambridge, MA, 1999. 4.1, 4.5

S. Edelman, K. Grill-Spector, T. Kushnir, and R. Malach. Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology*, 26(4): 309–321, 1998. 4.1, 4.5

J.V. Filoteo, W.T. Maddox, and J.D. Davis. Quantitative modeling of category learning in amnesic patients. *Journal of the International Neuropsychological Society*, 7(1): 1–19, 2001. 3.7.4

J. A. Fodor. Precis of the modularity of mind. *Behavioral and Brain Sciences*, 8(1): 1–5, 1985. 1.3

D.J. Freedman, M. Riesenhuber, T. Poggio, and E.K. Miller. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312–316, 2001. 2.1

I. Gauthier, A.W. Anderson, M.J. Tarr, P. Skudlarski, and J.C. Gore. Levels of categorization in visual recognition studied using functional magnetic resonance imaging. *Current Biology*, 7(9):645–651, 1997. 1.2

I. Gauthier, P. Skudlarsk, J.C. Gore, and A.W. Anderson. Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, 3(2): 191–197, 2000a. 1.2

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. *Nature Neuroscience*, 1999. 1.2

I. Gauthier, M.J. Tarr, J. Moylan, A.W. anderson, P. Skudlarski, and J.C. Gore. Does visual subordinate-level categorization engage the functionally-defined fusiform face area? *Journal of Cognitive Neuropsychology*, 2000b. 1.2

M.A. Goodale and A.D. Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992. 1.1

Z.M. Hafed and J.J. Clark. Microsaccades as an overt measure of covert attention shifts. *Vision Research*, 42(22):2533–2545, 2002. 6.1, 6.7

D.J. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9(2):181–197, 1992. 7.1

R.J. Hernstein and P.A. de Villiers. Fish as a natural category for people and pigeons. *The Psychology of Learning and Motivation*, 14:59–95, 1980. 2.2.3

J.E. Hoffman and B. Subramaniam. The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, 57(6):787–795, 1995. 6.1

L. Huber and R. Lenz. Categorization of prototypical stimulus classes by pigeons. *Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, 49(2):111–133, 1996. 2.2.1

L. Ingber. Very fast simulated re-annealing. *Mathematical and Computer Modelling*, 12(8):967–973, 1989. 3.4

A. Ishai, L.G. Ungerleider, A. Martin, H.L. Schouten, and J.V. Haxby. Distributed representation of objects in the human ventral visual pathway. *Proceedings of The National Academy of Sciences of the United States of America*, 96(16):9379–9384, 1999. 2.1

L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, Mar 2001. 6.1

L. Itti, C. Koch, and J. Braun. Revisiting spatial vision: Towards a unifying model. *Journal of the Optical Society of America, JOSA-A*, 17(11):1899–1917, Nov 2000. 7.2

L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (11):1254–1259, Nov 1998. 6.1, 6.5

W. James. *The Principles of Psychology*. Harvard University Press, Cambridge, Massachusetts, 1890. 6.1

J. Jovicich, R.J. Peters, C. Koch, C. Chang, and T. Ernst. Human perception of faces and face cartoons: an fMRI study. *Proc. 8th Scientific Meeting and Exhibition of the International Society for Magnetic Resonance in Medicine*, 2000. 2.2.2, 2.3

N. Kanwisher, J. McDermott, and M.M. Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302–4311, 1997. 1.2, 2.1, 2.2.2, 2.3

B.J. Knowlton. What can neuropsychology tell us about category learning? *Trends in Cognitive Sciences*, 3(4):123–124, 1999. 3.1, 3.7.4

B.J. Knowlton and L.R. Squire. The learning of categories—parallel brain systems for item memory and category knowledge. *Science*, 262(5140):1747–1749, 1993. 3.7.4

C. Koch. *The Quest for Consciousness: A Neurobiological Approach*. Roberts & Company, Denver, Colorado, 2004. 6.1

C. Koch and S. Ullman. Shifts in selective visual-attention—towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227, 1985. 6.5

E. Kowler, E. Anderson, B. Dosher, and E. Blaser. The role of attention in the programming of saccades. *Vision Research*, 35(13):1897–1916, 1995. 6.1

J.K. Kruschke. ALCOVE—an exemplar-based connectionist model of category learning. *Psychological Review*, 99(1):22–44, 1992. 3.1, 3.7.4

A.A. Kustov and D.L. Robinson. Shared neural control of attentional shifts and eye movements. *Nature*, 384(6604):74–77, 1996. 6.1, 6.7

D.K. Lee, L. Itti, C. Koch, and J. Braun. Attention activates winner-take-all competition among visual filters. *Nature Neuroscience*, 2(4):375–381, 1999. 7.1, 7.2, 7.1, 7.3, 9.3

W. Li and C.D. Gilbert. Global contour saliency and local colinear interactions. *Journal of Neurophysiology*, 88(5):2846–2856, 2002. 8.1

Z.P. Li. A neural model of contour integration in the primary visual cortex. *Neural Computation*, 10(4):903–940, 1998. 8.1

W.T. Maddox. On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics*, 61(2):354–374, 1999. 3.4

W.T. Maddox and F.G. Ashby. Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, 53(1):49–70, 1993. 2.1, 3.7

W.T. Maddox and F.G. Ashby. Selective attention and the formation of linear decision boundaries: Comment on McKinley and Nosofsky (1996). *Journal of Experimental Psychology: Human Perception and Performance*, 24(1):301–321, 1998. 3.7

D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman & Company, San Francisco, California, 1982. 6.5

S.C. McKinley and R.M. Nosofsky. Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception and Performance*, 22(2):294–317, 1996. 3.7

F. Miau and L. Itti. A neural model combining attentional orienting to object recognition: Preliminary explorations on the interplay between where and what. *Proc. IEEE Engineering in Medicine and Biology Society (EMBS), Istanbul, Turkey*, Oct 2001. 6.1

A.D. Milner and M.A. Goodale. Visual pathways to perception and action. *Progress iIn Brain Research*, 95:317–337, 1993. 1.1

M. Mishkin, L.G. Ungerleider, and K.A. Macko. Object vision and spatial vision—two cortical pathways. *Trends in Neurosciences*, 6(10):414–417, 1983. 1.1

T. Moore, K.M. Armstrong, and M. Fallah. Visuomotor origins of covert spatial attention. *Neuron*, 40(4):671–683, 2003. 6.1

T. Moore and M. Fallah. Control of eye movements and spatial attention. *Proceedings of The National Academy of Sciences of the United States of America*, 98(3): 1273–1276, 2001. 6.1

T. Moore and M. Fallah. Microstimulation of the frontal eye field and its effects on covert spatial attention. *Journal of Neurophysiology*, 91(1):152–162, 2004. 6.1

T.N. Mundhenk and L. Itti. A model of contour integration in early visual cortex. *Biologically Motivated Computer Vision, Proceedings*, 2525:80–89, 2002. 8.1, 8.2

J.A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965. 3.4

A.C. Nobre, D.R. Gitelman, E.C. Dias, and M.M. Mesulam. Covert visual spatial orienting and saccades: Overlapping neural systems. *Neuroimage*, 11(3):210–216, 2000. 6.1

R.M. Nosofsky. Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(1):104–114, 1984. 2.1, 3.1

R.M. Nosofsky. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39–57, 1986. 4.5

R.M. Nosofsky. Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, 34(4):393–418, 1990. 3.7

R.M. Nosofsky. Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, 17(1):3–27, 1991. 2.1, 2.2.1, 3.1, 3.2, 3.3.1, 3.7, 3.7.2, 1, 4.5

R.M. Nosofsky. Selective attention and the formation of linear decision boundaries: Reply to Maddox and Ashby (1998). *Journal of Experimental Psychology: Human Perception and Performance*, 24(1):322–339, 1998. 3.7

R.M. Nosofsky, J.K. Kruschke, and S.C. McKinley. Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2):211–233, 1992. 3.1, 3.7.5

H. Op de Beeck, J. Wagemans, and R. Vogels. Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience*, 4 (12):1244–1252, 2001. 2.1, 4.1, 4.5, 5.5

D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, 2002. (document), 6.1, 6.6, 6.7

D. Parkhurst and E. Niebur. What could over 1000 internet users tell us about visual attention? *Vision Sciences Society (VSS) Annual Meeting, Sarasota, Florida*, 2003. 9.3

M.W. Pettet and C.D. Gilbert. Dynamic changes in receptive-field size in cat primary visual-cortex. *Proceedings of The National Academy of Sciences of the United States of America*, 89(17):8366–8370, 1992. 8.1

T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247(4945):978–982, 1990. 3.1

U. Polat and D. Sagi. Lateral interactions between spatial channels—suppression and facilitation revealed by lateral masking experiments. *Vision Research*, 33(7): 993–999, 1993. 8.1

U. Polat and D. Sagi. The architecture of perceptual spatial interactions. *Vision Research*, 34(1):73–78, 1994a. 8.1

U. Polat and D. Sagi. Spatial interactions in human vision—from near to far via experience-dependent cascades of connections. *Proceedings of The National Academy of Sciences of the United States of America*, 91(4):1206–1209, 1994b. 8.1

M.I. Posner and Y. Cohen. Components of performance. In H. Bouma and D. Bowhuis, editors, *Attention and Performance X*, pages 531–556. Erlbaum, Hillsdale, NJ, 1984. 6.1

S.K. Reed. Pattern recognition and categorization. *Cognitive Psychology*, 3:382–407, 1972. 2.1, 2.2.1, 3.3.1, 3.3.4, 3.7

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999. 2.1, 3.7.4, 5, 5.1, 5.5, 5.2, 5.5

G. Rizzolatti, L. Riggio, I. Dascola, and C. Umilta. Reorienting attention across the horizontal and vertical meridians—evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25(1A):31–40, 1987. 6.1, 6.7

E. Rosch, C.B. Mervis, W.D. Gray, D.M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976. 1.2

Y. Rosseel. Connectionist models of categorization: A statistical interpretation. *Psychologica Belgica*, 36(1-2):93–112, 1996. 3.1

U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is attention useful for object recognition? *IEEE International Conference on Computer Vision and Pattern Recognition*, 2004. 6.1

G.E. Schneider. Two visual systems: brain mechanisms for localization and discrimination are dissociated by tectal and cortical lesions. *Science*, 163(3870):895–902, 1969. 1.1

B.M. Sheliga, L. Riggio, and G. Rizzolatti. Orienting of attention and eye movements. *Experimental Brain Research*, 98(3):507–522, 1994. 6.1, 6.7

B.M. Sheliga, L. Riggio, and G. Rizzolatti. Spatial attention and eye movements. *Experimental Brain Research*, 105(2):261–275, 1995. 6.1, 6.7

R.N. Shepard. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, 1987. 3.3.1, 4.5

N. Sigala, F. Gabbiani, and N.K. Logothetis. Visual categorization and object representation in monkeys and humans. *Journal of Cognitive Neuroscience*, 14(2):187–198, 2002. 2.2.1, 2.2.3, 3.2, 3.7, 3.7.4, 4.2.2, 1, 4.5

N. Sigala and N.K. Logothetis. Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, 415(6869):318–320, 2002. 2.1, 3.7.4, 4.1, 4.5, 5.5

V.M. Sloutsky and Y.F. Lo. How much does a shared name make things similar? Part 1: Linguistic labels and the development of similarity judgment. *Developmental Psychology*, 35(6):1478–1492, 1999. 4.5

J.D. Smith and J.P. Minda. Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, 24 (6):1411–1436, 1998. 2.1

L.R. Squire and B.J. Knowlton. Learning about categories in the absence of memory. *Proceedings of The National Academy of Sciences of the United States of America*, 92(26):12470–12474, 1995. 3.7.4

D.D. Stettler, A. Das, J. Bennett, and C.D. Gilbert. Lateral connectivity and contextual interactions in macaque primary visual cortex. *Neuron*, 36(4):739–750, 2002. 8.1

T. Sugihara, S. Edelman, and K. Tanaka. Representation of objective similarity among three-dimensional shapes in the monkey. *Biological Cybernetics*, 78(1):1–7, 1998. 2

J.W. Tanaka and M. Taylor. Object categories and expertise: Is the basic level in the eye of the beholder. *Cognitive Psychology*, 23:457–482, 1992. 1.2

Y. Tanaka and S. Shimojo. Location vs feature: Reaction time reveals dissociation between two visual functions. *Vision Research*, 36(14):2125–2140, 1996. 1.1

M.J. Tarr and I. Gauthier. Ffa: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience*, 3(8):764–769, 2000. 1.2

S. Treue. Visual attention: the where, what, how and why of saliency. *Current Opinion in Neurobiology*, 13(4):428–432, 2003. 6.1

L.G. Ungerleider and M. Mishkin. Two cortical visual systems. In D.J. Ingle, M.A. Goodale, and R.J.W. Mansfield, editors, *Analysis of Visual Behavior*, pages 549–586. MIT Press, Cambridge, Massachusetts, 1982. 1.1

J.H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of The Royal Society of London B*, 265:359–366, 1998. 6.2

R. Vogels. Categorization of complex visual images by rhesus monkeys. Part 1: behavioural study. *European Journal of Neuroscience*, 11(4):1223–1238, 1999. 2.2.3

D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional selection for object recognition—a gentle way. *Biologically Motivated Computer Vision—Lecture Notes in Compter Science*, 2525:472–479, 2002. 6.1

B.A. Wandell. *Foundations of Vision*. Sinauer Associates, Sunderland, Massachusetts, 1995. 6.1, 6.5

B. Zenger, J. Braun, and C. Koch. Attentional effects on contrast detection in the presence of surround masks. *Vision Research*, 40(27):3717–3724, 2000. 8.1

B. Zenger and D. Sagi. Isolating excitatory and inhibitory nonlinear spatial interactions involved in contrast detection. *Vision Research*, 36(16):2497–2513, 1996. 8.1

W. Zucchini. An introduction to model selection. *Journal of Mathematical Psychology*, 44(1):41–61, 2000. 3.4