Models of Bottom-Up and Top-Down Visual Attention

Thesis by

Laurent Itti

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

2000
(Submitted January 11, 2000)

Acknowledgements

My profound gratitude goes to Professor Christof Koch, my mentor and research advisor, and Dr. Jochen Braun and Prof. Ernst Niebur, who all contributed significantly to the work described in this thesis. Without their guidance, contributions of ideas and theories, and even contributions down to the details of mathematical derivations and software implementations, I would certainly not have been able to complete the more complex and interesting parts of this work. I salute Professor Koch's rich ideas, genuine passion for science, discovery and understanding, superior mentoring skills, and unparalleled availability, with a maximum time-to-meeting of approximately 15 minutes in any emergency. I also salute Dr. Braun's bright scientific ideas, solid methods, and patience with respect to my often lighter approach to research. Finally, Professor Niebur's sharp scientific mind, enthusiasm and efficient thinking have played a key role in getting me started on this work.

My very special thanks go to my family, which, despite the great distance, have always given me love and unconditional support. I also would like to thank Professors Linda Chang and Thomas Ernst, at the Harbor-UCLA Medical Center, for their friendship as well as the opportunity to participate to their amazingly productive research.

I am indebted to the members of my thesis committee, Professors Pietro Perona, Richard Andersen, Demetri Psaltis and Shinsuke Shimojo, for their exceptional receptivity and efficient advice each time I seeked their help. In particular, Professor Perona contributed a number of ideas to the work presented here, and kindly provided laboratory space and computer resources for our psychophysical and computational experiments.

A number of persons have been kind enough to provide me with experimental data or databases of images used in this thesis. In particular, Dr. Dale K. Lee provided us with the psychophysical data on attentional modulation of spatial vision thresholds used with the second model presented here. The "Traffic Signs" images used with our first model were kindly provided to us by Werner Ritter at Daimler-Benz. The "Military Vehicle" database was sent to us by Dr. Lex Toet at TNO-Human Factors Research Institute, The Netherlands. The "Pedestrians" database was kindly made available to us by Constantine Papageorgiou, who works with Professor Tommy Poggio at MIT/CBCL. The "Good Housekeeping" magazine covers were submitted to us by Peter Walker, Director of the Neural Nets Group at the McCann-Erickson advertising agency. Finally, the preliminary eye movement data was acquired in collaboration with Drs. Steffen Egner and Christian Scheier.

Throughout this research, a number of people have contributed key ideas and criticisms during the development of our models. In particular, my friend and office mate, Amit Manwani, has rarely known a day when I did not submit to him my current problems or concerns. The many others include Dr. Ching E. Ho, Prof. Alex Pouget, Dr. Dr. Peter Steinmetz, Grace Chang, Dr. Geraint Rees, Dr. Barbara Zenger, and all whom I forget.

I would not have found the energy to work hard without the support and encouragements of all my friends, who will recognize themselves.

This work was supported by ONR, NSF and NIHM.

Abstract

When we observe our visual environment, we do not perceive all its components as being equally interesting. Some objects automatically and effortlessly "pop-out" from their surroundings, that is, they draw our visual attention, in a **"bottom-up"** manner, towards them. In a first approximation, focal visual attention acts as a rapidly shiftable "spotlight," which allows only the selected information to reach higher levels of processing and representation. Most models of the bottom-up control of attention are based on the concept of a *saliency map*, that is, an explicit two-dimensional map that encodes the conspicuity of objects in the visual environment. Competition among neurons in this map gives rise to a single winning location that corresponds to the next attended target. Inhibiting this location automatically allows the system to attend to the next most salient location. A first body of work in this thesis describes a detailed computer implementation of such a scheme, focusing on the problem of combining information across modalities, here orientation, intensity and color information, in a purely stimulus-driven manner. The model is applied to common psychophysical stimuli as well as to very demanding visual search tasks. Its successful performance is used to address the extent to which the primate visual system carries out visual search via one or more such saliency maps and how this can be tested.

We next address the question of what happens once our attention is focused onto a restricted part of our visual field. There is mounting experimental evidence that attention is far more sophisticated than a simple feed-forward spatially-selective filtering process. Indeed, visual processing appears to be significantly different inside the attentional spotlight than outside. That is, in addition to its properties as a feed-forward information processing and transmission bottleneck, focal visual attention feeds back and locally modulates, in a **"top-down"** manner, the visual processing and representation of selected objects. The second body of work presented in this thesis is concerned with a detailed computational model of basic pattern vision in humans and its modulation by top-down attention. We start by acquiring a complete dataset of five different simple psychophysical experiments, including discriminations of contrast, orientation and spatial frequency of simple pattern stimuli by human observers. This experimental dataset places strict constraints on our model of early pattern vision. The model, however, is eventually able to reproduce the entire dataset while assuming plausible neurobiological components. The model is further applied to existing psychophysical data which demonstrates how top-down attention alters performance in these simple psychophysical discrimination experiments. Our model is able to quantitatively account for all observations by assuming that attention strengthens the non-linear cortical interactions among visual neurons.

Together, the two aspects of attention studied in this thesis lead us to consider the essential role of non-linear computations in visual processing. We suggest that visual processing, even at its earliest levels, is best characterized not by linear response functions and spatial convolutions, but rather by non-linearly interacting computational devices.

Contents

List of Figures

List of Tables

Part I

General Introduction

Selecting only a subset of the available sensory information before further detailed processing is crucial for efficient perception. In the visual modality, this selection is frequently implemented by suppressing information outside a spatially circumscribed region of the visual field, the so-called "focus of attention." In primates, the identification of behaviorally relevant objects and the analysis of their spatial relationships involves either rapid, saccadic eye movements or so-called "covert" (i.e., without eye movements) shifts of visual attention. In a first approximation, focal attention acts as a rapidly shiftable "spotlight," which allows the selected information to enter short-term memory and to remain there long enough to reach conscious and cognitive levels of representation. By employing such spatially serialized analysis strategy, primate brains achieve unparalleled performance, given the limited computational resources available, in scene interpretation and understanding. While attention can be controlled in a voluntary manner, it is also attracted in a **"bottom-up,"** automatic and unconscious manner to conspicuous, or "salient," visual locations. This last property is of particular behavioral importance, because it constitutes a powerful alerting system which allows primates to instantly become aware of unexpected predators. The first part of this thesis is concerned with a detailed and neurobiologically-plausible model of the bottom-up control of visual attention.

What then happens when we focus our attention to a restricted part of our visual field? As we just mentioned, focal attention is often thought as a gating mechanism, which selectively allows a certain spatial location and certain types of visual features to reach higher visual processes. The question then naturally arises of the nature, computational characteristics and neurobiological implementation of such information processing bottleneck. Experimental work, in particular psychophysical experiments in humans and electrophysiological experiments in the awake behaving monkey, suggest that the attentional bottleneck is not unidirectional. Indeed, it has been observed in our and other laboratories that focal attention modulates behavioral performance in simple pattern discrimination tasks, a finding which is paralleled by observations of a **"top-down"** modulation of neuronal activity by attention. Thus, it appears that early visual processing is significantly different inside than outside of the attentional spotlight. The second part of this thesis describes a detailed biological model of early spatial vision and how such model provides a quantitative computational account of the observed attentional modulation of early visual processing.

The present document is organized as follows. In **Part II**, we present a simple computational model which attempts to mimic the low-level, automatic and unconscious neuronal mechanisms responsible for attracting our attention to salient objects in our environment. After a brief introduction and review of existing modeling work on saliency-based attentional guidance in **Chapter 1**, we extensively describe the proposed model in **Chapter 2**. Our "bottom-up" model closely follows the neuronal architecture of the earliest hierarchical levels of visual processing. The retinal input image is first decomposed into a multiscale representation, and further into a set of topographic maps selective for particular visual attributes. Such "feature maps" are composed of visually responsive neurons whose responses already are much more sophisticated than the pixel sensors of electronic cameras: For example, such neurons may maximally respond to oriented edges, isolated bright spots on dark surrounds, local motion contrast, or several types of local color contrasts. In addition, complex cortical interactions further shape the response of these neurons, in particular by enforcing a strong spatial competition for activity. The outputs of all feature maps are combined into a unique "saliency map," which encodes for local conspicuity in a more abstract manner, independent of the particular visual features which render an object salient. At any given time, a so-called "winner-take-all" neural network selects the most active location in the saliency map and draws the attentional spotlight towards it. Subsequently, the selected object is suppressed in the saliency map, such that the winner-take-all will select the next most salient location. With time and in the absence of cognitive or volitional influences, the system generates attentional scanpaths which indicate which locations were selected as being of interest to the model. In **Chapter 3** we study a number of applications of our model, ranging from its successful reproduction of simple visual search experiments, to systematic studies of its performance as a target detection system when analyzing complex natural scenes. Finally, in **Chapter 4**, we review our computational results within the broader frameworks of biological and artificial vision systems, and propose further directions of research.

Part III is devoted to experiments and modeling on human pattern vision, and to a detailed computational account of how "top-down" attention modulates early vision. This Part consequently directly follows and complements **Part II**, by studying in detail what happens once attention is focused onto a particular region of the visual scene.

After a brief introduction in **Chapter 5** to the experimental and computational methods to be used, we start by describing experiments in **Chapter 6** and modeling in **Chapter 7** which are aimed at deriving a simple, unifying model for early pattern vision in humans. We use psychophysical experiments to constrain and validate our model until it accurately reproduces human performance. Indeed, human psychophysics, an experimental method which aims at inferring the computational architecture of the visual system by measuring observer responses to simple visual stimuli, has emerged as one of the most advanced and well characterized experimental methods in humans. Psychophysical thresholds for stimulus contrast, orientation, and spatial frequency have been studied for several decades. Quantitative accounts of these thresholds have become increasingly refined and usually involve a population of "noisy filters" tuned to different orientations and spatial frequencies. Although earlier models postulated filters that are independent of each other, there are serious shortcomings to this approach. More recent models postulate an interaction between filters with spatially overlapping receptive fields, specifically, the normalization of individual filter responses relative to the total response of the local filter population ("divisive inhibition"). An intriguing parallel to these perceptual accounts can be found in certain models of visual cortical responses to stimulus contrast and orientation. Despite marked differences in detail, the models in question consider a population of neurons with overlapping receptive fields, broadly tuned to a range of different orientations, and normalize individual responses relative to the population response. Our experiments are directly targeted towards a detailed quantitative characterization for such interactions, by using spatially localized stimuli presented in the near-peripheral visual field. Acquiring a consistent dataset, in three observers, consisting of five simple experiments, allows us to derive strong computational constraints on early visual processing. Closely related to previously proposed models of basic pattern vision, we thus formulate our model as a "consensus" or "unifying" neuronal model, which simultaneously accounts in a precise quantitative manner for the broad range of experimental results acquired in our laboratory, while assuming neuronally plausible components.

While this model is developed to account for vision when attention is fully available to the pattern discrimination task studied, we then apply it in **Chapter 8** to human data by Lee (Lee, 1999), which shows how top-down attention has a modulatory influence on early visual processes. Indeed, it has been extensively established using a number of experimental methods in primates that shifting attention away from a visual stimulus reduces, but does not abolish, visual discrimination performance. This residual vision with "poor" attention can be compared to normal vision with "full" attention to reveal how attention alters visual perception. Large differences have been reported between residual and normal visual thresholds for discriminating the orientation or spatial frequency of simple patterns, and smaller differences for discriminating contrast. We specifically investigate the meaning of this observed attentional modulation of visual thresholds using our unifying model of spatial vision. Our model quantitatively accounts for all observations, and predicts that the effects of attention on visual cortical neurons include increased contrast gain as well as sharper tuning to orientation and spatial frequency. Together, these two effects suggest that attention activates a winner-take-all competition amongst overlapping visual filters.

Together, our models of bottom-up and of top-down attention have taught us a number of simple lessons on basic computational principles in early primate vision. Our results and findings are summarized in **Part IV**. In particular, we put our experience in perspective within the broader concept of neuronal receptive fields and early vision. We suggest that visual processing, even at its earliest hierarchical levels, is best characterized not by linear response functions and spatial convolutions, but rather by highly non-linear and interacting localized computational devices.

Part II

A Model of Bottom-up, Saliency-based Attention

Chapter 1 Introduction

Most biological vision systems (including *Drosophila*; (Heisenberg & Wolf, 1984) appear to employ a serial computational strategy when inspecting complex visual scenes. Particular locations in the scene are selected based on their behavioral relevance or on local image cues. In primates, the identification of objects and the analysis of their spatial relationship usually involves either rapid, saccadic eye movements to bring the fovea onto the object, or covert shifts of attention.

It may seem ironic that brains employ serial processing, since one usually thinks of them as paradigmatic "massively parallel" computational structures. However, in any physical computational system, processing resources are limited, which leads to bottlenecks similar to those faced by the *von Neumann* architecture on conventional digital machines. Nowhere is this more evident than in the primate's visual system, where the amount of information coming down the optic nerve—estimated to be on the order of of 10^8 bits per second—far exceeds what the brain is capable of fully processing and assimilating into conscious experience. The strategy nature has devised for dealing with this bottleneck is to select certain portions of the input to be processed preferentially, shifting the processing focus from one location to another in a serial fashion.

Despite the widely shared belief in the general public that "we see everything around us," only a small fraction of the information registered by the visual system at any given time reaches levels of processing that directly influence behavior. This is vividly demonstrated by *change blindness* (O'Regan *et al.*, 1999; Simons & Levin, 1997) in which significant image changes remain nearly invisible under natural viewing conditions, although observers demonstrate no difficulty in perceiving these changes once directed to them. Overt and covert attention controls access to these privileged levels and ensures that the selected information is relevant to behavioral priorities and objectives. Operationally, information can be said to be "attended" if it enters short-term memory and remains there long enough to be voluntarily reported. Thus, visual attention is closely linked to *visual awareness* (Crick & Koch, 1998).

But how is the selection of one particular spatial location accomplished? Does it involve primarily bottom-up, sensory-driven cues or does expectation of the targets characteristics play a decisive role? A large body of literature has concerned itself with the psychophysics of visual search or orienting for targets in sparse arrays or in natural scenes using either covert or overt shifts of attention (for reviews, see (Niebur & Koch, 1998) or the survey article (Toet *et al.*, 1998)).

1.1 Two-Component Framework of Attention

Much evidence has accumulated in favor of a two-component framework for the control of where in a visual scene attention is deployed (James, 1890/1981; Treisman & Gelade, 1980;

Bergen & Julesz, 1983; Treisman, 1988; Nakayama & Mackeben, 1989; Braun & Sagi, 1990; Hikosaka *et al.*, 1996; Braun & Julesz, 1998; Braun, 1998b): A bottom-up, fast, primitive mechanism that biases the observer towards selecting stimuli based on their *saliency* (most likely encoded in terms of center-surround mechanisms) and a second slower, top-down mechanism with variable selection criteria, which directs the "spotlight of attention" under cognitive, volitional control. Whether visual consciousness can be reached by either saliency-based or top-down attentional selection or by both remains controversial.

Preattentive, parallel levels of processing do not represent all parts of a visual scene equally well, but instead provide a weighted representation with strong responses to a few parts of the scene and poor responses to everything else. Indeed, in an awake monkey freely viewing a natural visual scene, there are not many locations which elicit responses in visual cortex comparable to those observed with isolated, laboratory stimuli (Gallant *et al.*, 1998). Whether a given part of the scene elicits a strong or a poor response is thought to depend very much on "context," that is, on what stimuli are present in other parts of the visual field. In particular, the recently accumulated evidence for "non-classical" modulation of a cell's response by the presence of stimuli outside of the cell's receptive field provides direct support for the idea that different visual locations compete for activity (Sillito *et al.*, 1995; Sillito & Jones, 1996; Levitt & Lund, 1997). Those parts which elicit a strong response are thought to draw visual attention to themselves and to therefore be experienced as "visually salient." Directing attention at any of the other parts is thought to require voluntary "effort."

Both modes of attention can operate at the same time and visual stimuli have two ways of penetrating to higher levels of awareness: Being willfully brought into the focus of attention, or winning the competition for saliency.

1.2 Bottom-Up Attention and the Saliency Map

Koch and Ullman (Koch & Ullman, 1985) introduced the idea of a saliency map to accomplish preattentive selection (see also the concept of a "master map" in (Treisman, 1988)). This is an explicit two-dimensional map that encodes the saliency of objects in the visual environment. Competition among neurons in this map gives rise to a single winning location that corresponds to the most salient object, which constitutes the next target. If this location is subsequently inhibited, the system automatically shifts to the next most salient location, endowing the search process with internal dynamics.

Many computational models of human visual search have embraced the idea of a saliency map under different guises (Treisman, 1988; Wolfe, 1994; Itti *et al.*, 1998b; Niebur & Koch, 1996; Olshausen *et al.*, 1993). The appeal of an explicit saliency map is the relatively straightforward manner in which it allows the input from multiple, quasi-independent feature maps to be combined and to give rise to a single output: The next location to be attended. Electrophysiological evidence points to the existence of several neuronal maps, in the pulvinar, the superior colliculus and the intraparietal sulcus, which appear to specifically encode for the saliency of a visual stimulus (Robinson & Petersen, 1992; Rockland *et al.*, 1999; Gottlieb *et al.*, 1998; Colby & Goldberg, 1999).

However, some researchers reject the idea of a topographic map in the brain whose *raison d'être* is the representation of salient stimuli. In particular, Desimone and Duncan (Desimone & Duncan, 1995) postulate that selective attention is a consequence of interactions among feature maps, each of which encodes in an implicit fashion, the saliency of a stimulus in that particular feature. We know of only a single implementation of this idea in terms of a computer algorithm (Hamker, 1999).

We here describe a computer implementation of a preattentive selection mechanism based on the architecture of the primate visual system. We address the thorny problem of how information from different modalities—in the case treated here from 42 maps encoding intensity, orientation and color in a center-surround fashion at a number of spatial scales—can be combined into a single saliency map. In the next Chapter, we show how our algorithm qualitatively reproduces human performance on a number of classical search experiments, and evaluate it with a number of real scenes.

Chapter 2 Architecture of the Proposed Model

2.1 Overview

The present model is limited to the bottom-up control of attention, *i.e.*, to the control of selective attention by the properties of the visual stimulus. It does not incorporate any top-down, volitional component. Furthermore, we are here only concerned with the localization of the stimuli to be attended ("Where"), not their identification ("What"). A number of authors (Olshausen *et al.*, 1993; Beymer & Poggio, 1996) have presented models for the neuronal expression of attention along the occipital-temporal pathway once spatial selection has occurred.

We make the following four assumptions: First, visual input is represented, in early visual structures, in the form of iconic (appearance-based) topographic feature maps. Two crucial steps in the construction of these representations consist of center-surround computations in every feature at different spatial scales, and within-feature spatial competition for activity. Second, information from these feature maps is combined into a single map which represents the local "saliency" of any one location with respect to its neighborhood. Third, the maximum of this *saliency map* is, by definition, the most salient location at a given time, and it determines the next location of the attentional searchlight. And fourth, the saliency map is endowed with internal dynamics allowing the perceptive system to scan the visual input such that its different parts are visited by the focus of attention in the order of decreasing saliency.

Fig. 2.1. shows an overview of Koch and Ullman's (1985) and of our model. Input is provided in the form of digitized images, from a variety of sources including a consumer-electronics NTSC video camera.

2.2 Extraction of Early Visual Features

Given an input image, the first processing step consists of decomposing this input into a set of distinct "channels," by using linear filters tuned to specific stimulus dimensions, such as luminance, red, green, blue and yellow hues, or various local orientations. The number and response properties of the linear filters implemented in the model have been chosen according to what is known of their neuronal equivalents in the early stages of visual processing in primates (see below). In addition, such decomposition is performed at a number of spatial scales, to allow the model to represent smaller and larger objects in separate subdivisions of these channels.

2.2.1 Pyramidal Representation

Different spatial scales are created using Gaussian pyramids (Burt & Adelson, 1983), which consist of progressively low-pass filtering and sub-sampling the input image. In our im-

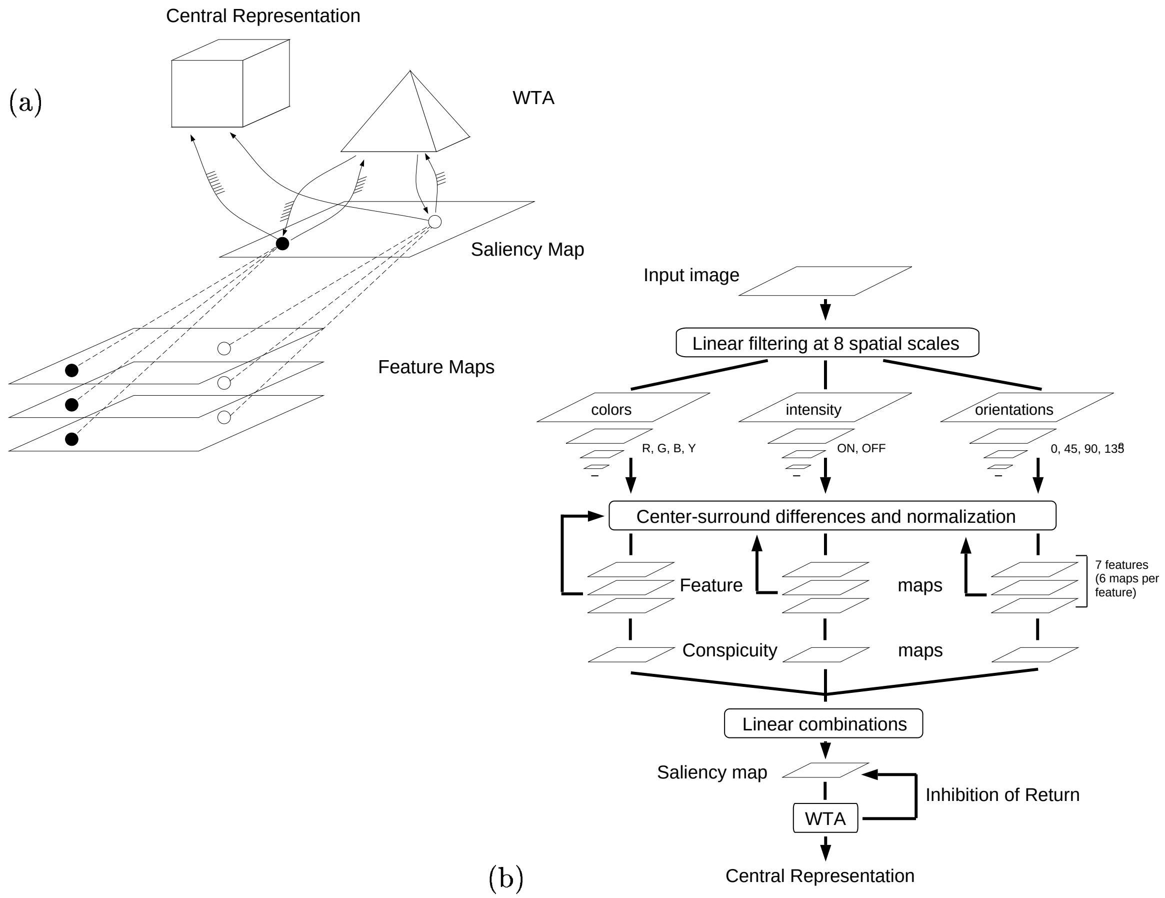


Figure 2.1: Schematic representation of Koch and Ullman's (1985) and of our model. **(a)** Original model of saliency-based visual attention, adapted from Koch and Ullman (1985). Early visual features such as color, intensity or orientation are computed, in a massively parallel manner, in a set of pre-attentive feature maps based on retinal input (not shown). Activity from all feature maps is combined at each location, giving rise to activity in the topographic saliency map. The winner-take-all (WTA) network detects the most salient location and directs attention towards it, such that only features from this location reach a more central representation for further analysis. **(b)** Schematic diagram for the model used in this study. It directly builds on the architecture proposed in **(a)**, but provides a complete implementation of all processing stages. Visual features are computed using linear filtering at eight spatial scales, followed by center-surround differences, which compute local spatial contrast in each feature dimension for a total of 42 maps. An iterative lateral inhibition scheme instantiates competition for salience within each feature map. After competition, feature maps are combined into a single "conspicuity map" for each feature type. The seven conspicuity maps then are summed into the unique topographic saliency map. The saliency map is implemented as a 2-D sheet of Integrate-and-Fire (I&F) neurons. The WTA, also implemented using I&F neurons, detects the most salient location and directs attention towards it. An Inhibition-of-Return mechanism transiently suppresses this location in the saliency map, such that attention is autonomously directed to the next most salient image location. We here do not consider the computations necessary to identify a particular object at the attended location.

plementation, pyramids have a depth of 9 scales, providing horizontal and vertical image reduction factors ranging from 1:1 (level 0; the original input image) to 1:256 (level 8) in consecutive powers of two **(Fig. 2.2)**. A 5×5 Gaussian filter is applied to each level of the pyramid before the decimation operation which yields the next level. Because successive convolutions by Gaussian filters are equivalent to a single convolution by a Gaussian filter (with a width which can be computed formally), pixels at various levels of the pyramid represent the responses of Gaussian filters with increasing width; in our case, the Gaussian width of a pixel at scale $\sigma + 1$ is equal to $\sqrt{5}$ times the width at scale σ.

With r, g and b being the red, green and blue channels of the input image, an intensity image I is obtained as $I = (r + g + b)/3$. I is used to create a Gaussian pyramid $I(\sigma)$, where $\sigma \in [0..8]$ is the scale. While more accurate expressions exist for the computation of luminance from a chromatic image, usually these expressions are dependent on a particular color acquisition, encoding or representation system (Foley *et al.*, 1990). For example, luminance is defined in the YIQ system (used for commercial television broadcasting in the U.S.A., using the NTSC encoding) as $0.299r + 0.587g + 0.114b$ (Foley *et al.*, 1990). Since our model was designed to be used with images from a variety of sources, we have not considered such more sophisticated recipes. It should also be pointed out that the luminous efficiency function for the human eye, *i.e.*, the eye's response to light of constant luminance, peaks at a wavelength of about 550nm, corresponding to yellow-green light (Foley *et al.*, 1990). This should be taken into account when trying to more closely reproduce the response properties of luminance-selective neurons.

The r, g and b channels are normalized by I in order to decouple hue from intensity. However, because hue variations are not perceivable at very low luminance (and hence are not salient), normalization is only applied at the locations where I is larger than 1/10 of its maximum over the entire image (other locations yield zero r, g and b). Four broadly-tuned color channels are created: $R = r - (g + b)/2$ for red, $G = g - (r + b)/2$ for green, $B = b - (r + g)/2$ for blue, and $Y = r + g - 2(|r - g| + b)$ for yellow (negative values are set to zero). Each channel yields maximal response for the pure, fully-saturated hue to which it is tuned, and yields zero response both for black and for white inputs. Four Gaussian pyramids $R(\sigma), G(\sigma), B(\sigma)$ and $Y(\sigma)$ are created from these color channels. **Fig. 2.3** shows a rendition of the chromatic selectivity of the four color-sensitive channels implemented in the model. This representation, like in the case of luminance described above, is very crude and only approximate. For instance, it does not account for the fact that the three types of color-sensitive "cone" photoreceptors in the human retina have their peak sensitivities at wavelengths of light which are not necessarily matched to the primary colors used in digitized images. (Although these three types of cones are often referred to as "red," "green" and "blue" types, with peak sensitivities at 580nm, 545nm and 440nm, they actually maximally respond to orange, yellow and blue hues, respectively; also, sensitivity to blue is almost ten

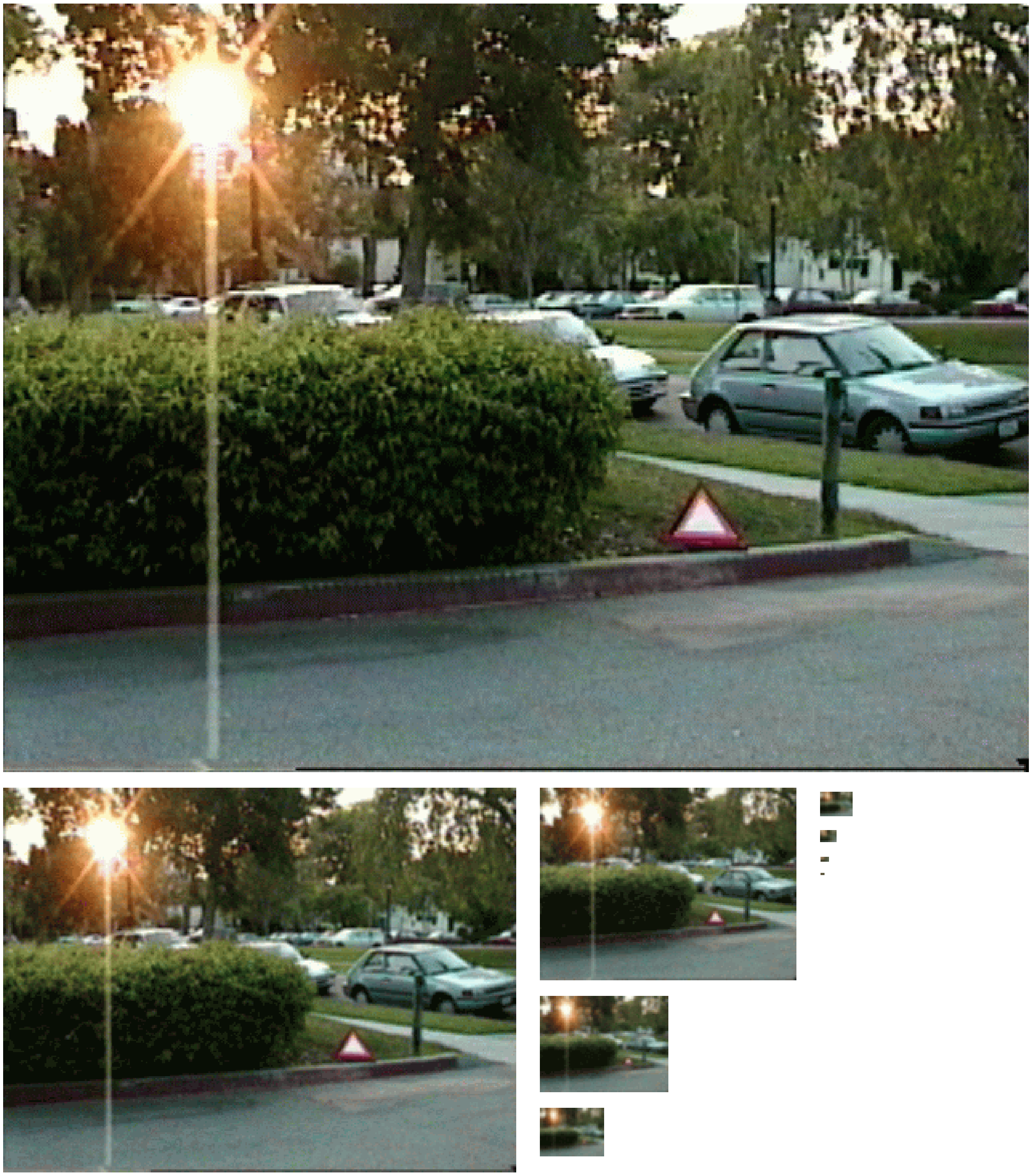


Figure 2.2: Pyramidal representation of the input image. The original image is shown on top, and rests at level 0 in the pyramid. Each subsequent level is obtained by low-pass filtering the previous level and downscaling it by a factor 2 in the horizontal and vertical directions.

times smaller than to red or green (Foley *et al.*, 1990).) The very simple expressions used here to compute R, G, B and Y are hence very simplified and approximate, which is mostly justified by the fact that they are also very simple and fast to compute.

Local orientation information is obtained from I using oriented Gabor pyramids $O(\sigma, \theta)$, where $\sigma \in [0..8]$ represents the scale and $\theta \in \{0°, 45°, 90°, 135°\}$ is the preferred orientation. (Gabor filters, which are the product of a cosine grating and a 2D Gaussian envelope, approximate the receptive field sensitivity profile (impulse response) of orientation-selective neurons in primary visual cortex (Leventhal, 1991).) The fast implementation proposed by Greenspan *et al.* is used in our model (Greenspan *et al.*, Jun. 1994). Some of the oriented filters and their responses on a test image are shown in **Fig. 2.4**.

2.2.2 Center-Surround Receptive Field Profiles

Each feature is computed in a center-surround structure akin to visual receptive fields. Using this biological paradigm renders the system sensitive to local spatial contrast in a given feature rather than to amplitude in that channel.

Center-surround operations are implemented in the model as differences between a fine and a coarse scale for a given feature: The center of the receptive field corresponds to a pixel at level $c \in \{2, 3, 4\}$ in the pyramid, and the surround to the corresponding pixel at level $s = c + \delta$, with $\delta \in \{3, 4\}$. We hence compute six feature maps for each type of feature (at scales 2-5, 2-6, 3-6, 3-7, 4-7, 4-8; see **Fig. 2.5**). Across-scale difference between two maps, denoted "$\ominus$" below, is obtained by interpolation to the finer scale and point-by-point subtraction. Using several scales not only for c, but also for $\delta = s - c$, yields truly multiscale feature extraction, by including different size ratios between the center and surround regions (contrary to previously used fixed ratios (Milanese *et al.*, 1995)). Seven types of features, for which wide evidence exists in mammalian visual systems, are computed in this manner from the low-level pyramids: As detailed below, one feature type encodes for on/off image intensity contrast (Leventhal, 1991), two encode for red/green and blue/yellow double-opponent channels (Luschow & Nothdurft, 1993; Engel *et al.*, 1997), and four encode for local orientation contrast (DeValois *et al.*, 1982; Tootell *et al.*, 1988).

2.2.3 The Feature Maps

The six feature maps for the intensity feature type encode for the modulus of image luminance contrast, i.e., the absolute value of the difference between intensity at the center (one of the three c scales) and intensity in the surround (one of the six $s = c + \delta$ scales). Center-surround differences ($\ominus$ defined previously) between a "center" fine scale c and a "surround" coarser scale s yield the feature maps. The first set of feature maps is concerned with intensity contrast, which in mammals is detected by neurons sensitive either to dark

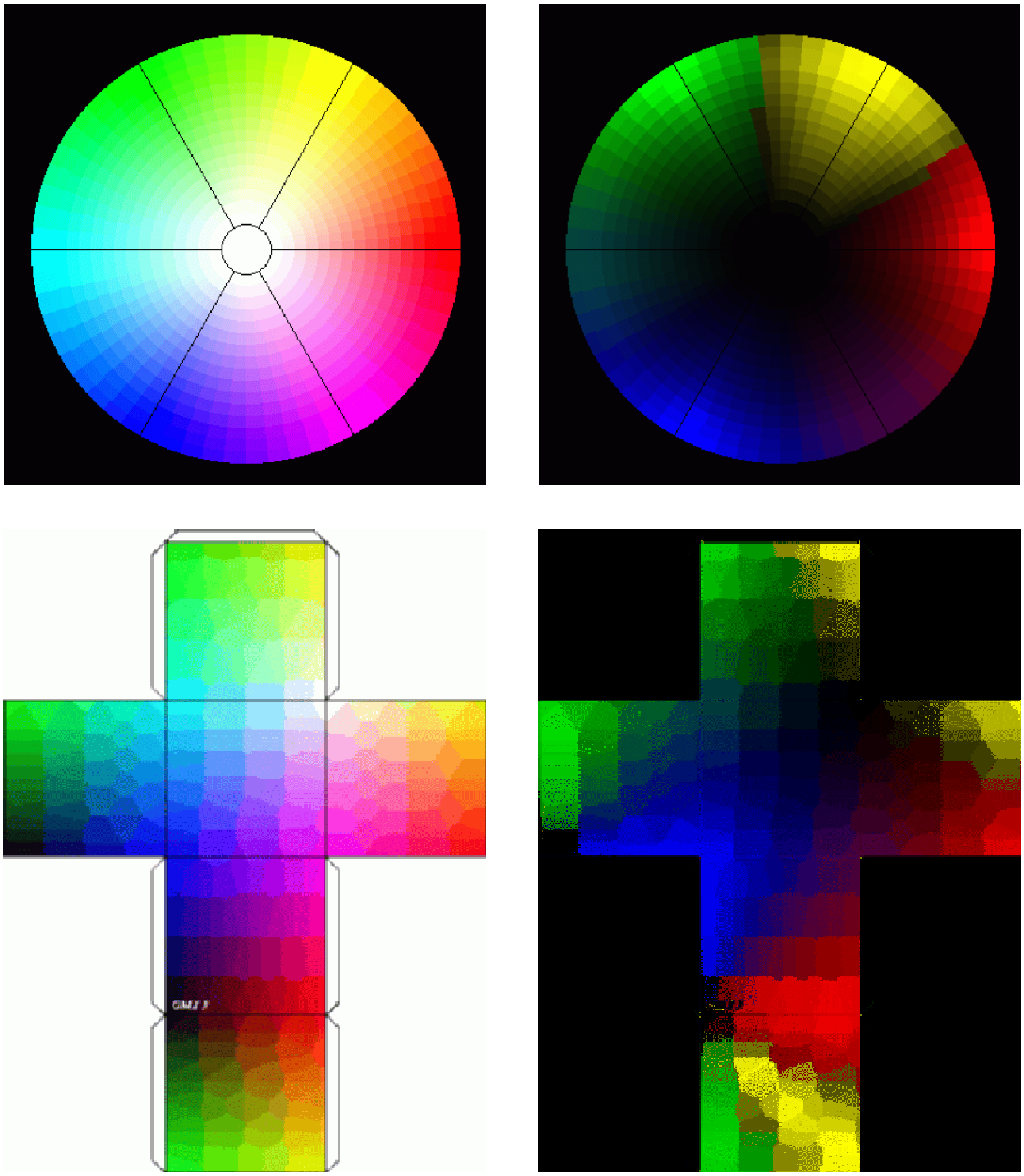


Figure 2.3: Color-selective channels implemented in the model, which emphasize chrominance information while attenuating the influence of luminance. **Left:** Original test images. **Right:** Composite renditions of the responses from the four luminance-normalized color channels R, G, B and Y.

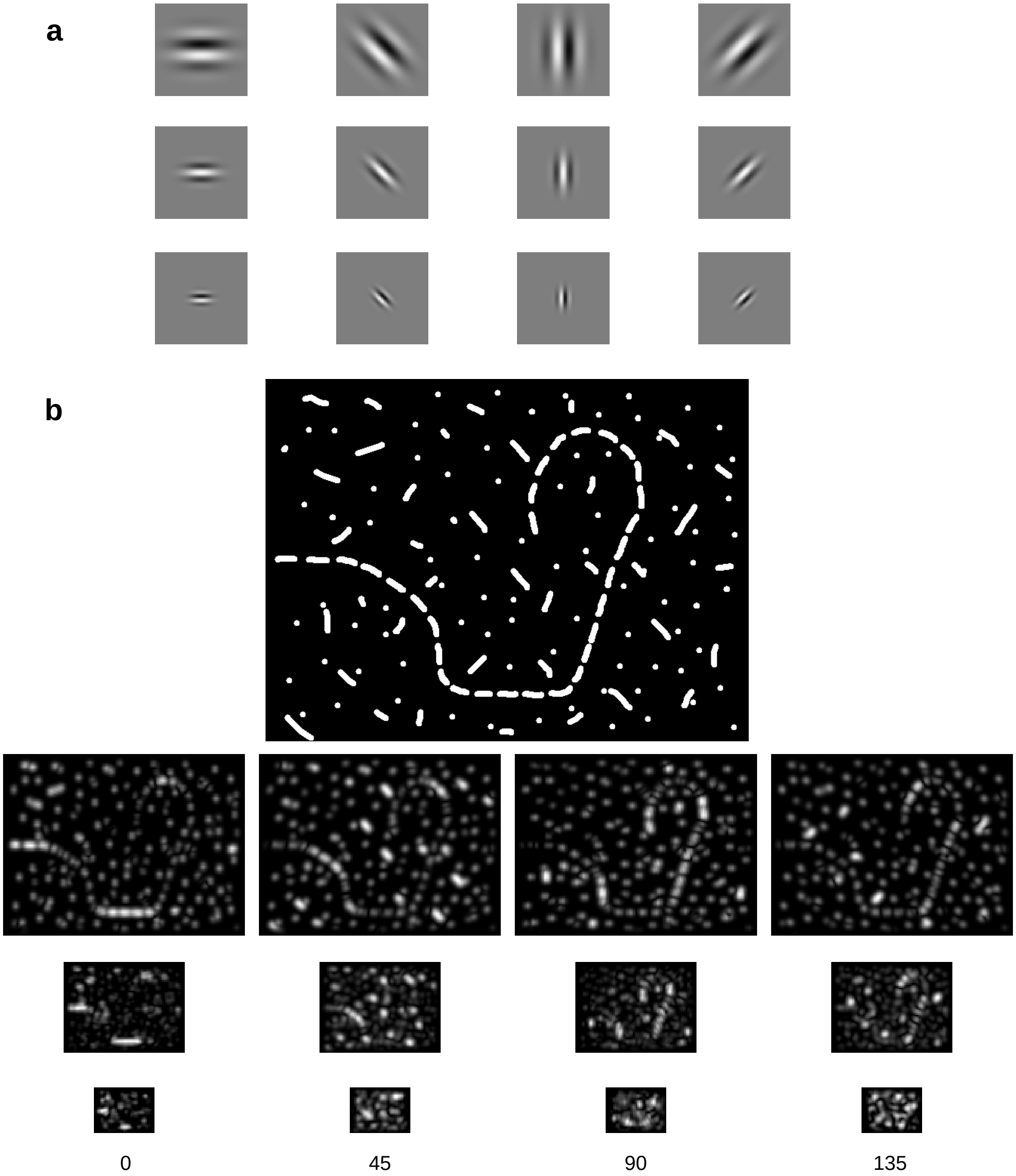
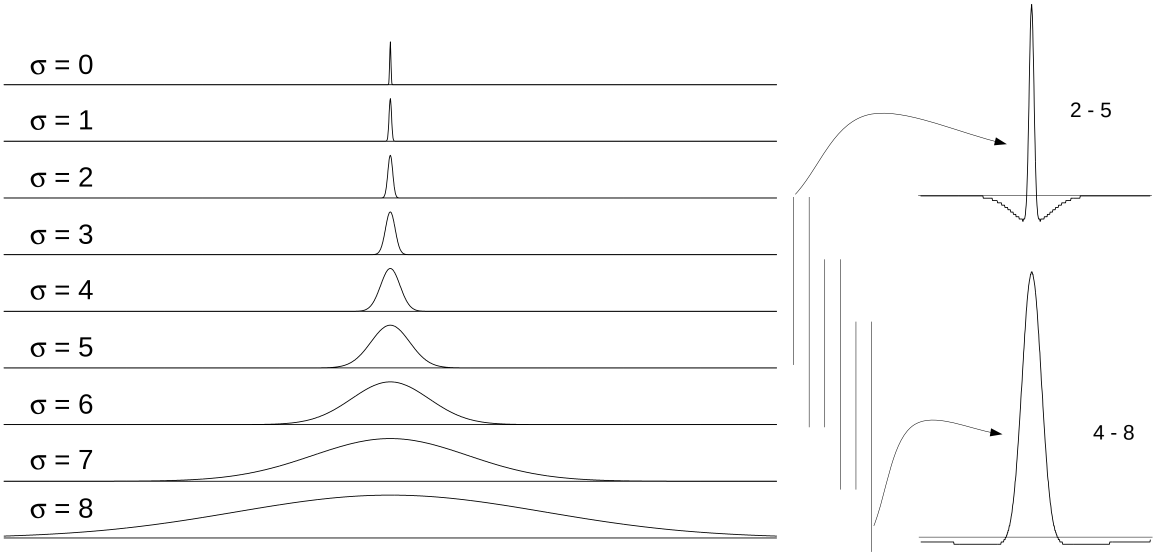


Figure 2.4: Orientation-selective Gabor filters. **(a)** Grey-scale renditions of the filters for three of the spatial scales and the four orientations implemented. **(b)** A test image and the responses from the filters shown in **(a)**.


σ = 0
σ = 1
σ = 2
σ = 3
σ = 4
σ = 5
σ = 6
σ = 7
σ = 8
2 - 5
4 - 8

centers on bright surrounds, or to bright centers on dark surrounds (Leventhal, 1991). Here, both types of sensitivities are simultaneously computed (using a rectification) in a set of six maps $\mathcal{I}(c, s)$, with $c \in \{2, 3, 4\}$ and $s = c + \delta$, $\delta \in \{3, 4\}$:

$$\mathcal{I}(c, s) = |I(c) \ominus I(s)| \tag{2.1}$$

A quantity corresponding to the double-opponency cells in primary visual cortex is then computed by center-surround differences across the normalized color channels. Each of the six red/green feature maps is created by first computing (red-green) at the center, then subtracting (green-red) from the surround, and finally outputting the absolute value. Accordingly, maps $\mathcal{RG}(c, s)$ are created in the model to simultaneously account for red/green and green/red double opponency (**Eq. 2.2**), and $\mathcal{BY}(c, s)$ for blue/yellow and yellow/blue double opponency (**Eq. 2.3**):

$$\mathcal{RG}(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))| \tag{2.2}$$

$$\mathcal{BY}(c, s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))| \tag{2.3}$$

The orientation feature maps are obtained from absolute center-surround differences between the orientation-selective channels. These maps, $\mathcal{O}(c, s, \theta)$, encode, as a group, local orientation contrast between the center and surround scales:

$$\mathcal{O}(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)| \tag{2.4}$$

It is possible in our implementation to use an arbitrary number of orientations, but we noticed that using more oriented filters than the four abovementioned did not alter the performances of the model drastically.

In total, 42 feature maps are thus computed: Six for intensity, 12 for color and 24 for orientation. **Fig. 2.6** shows, for an example image, all the feature and conspicuity maps described in this Section.

2.3 Combining Information Across Multiple Maps

Our modeling hypotheses assume the existence of a unique topographic saliency map. The purpose of the saliency map is to represent the conspicuity — or "saliency" — at every location in the visual field by a scalar quantity, and to guide the selection of attended locations, based on the spatial distribution of saliency. A combination of the feature maps provides bottom-up input to the saliency map, modeled as a dynamical neural network. At each spatial location, activity from the 42 feature maps consequently needs to be combined

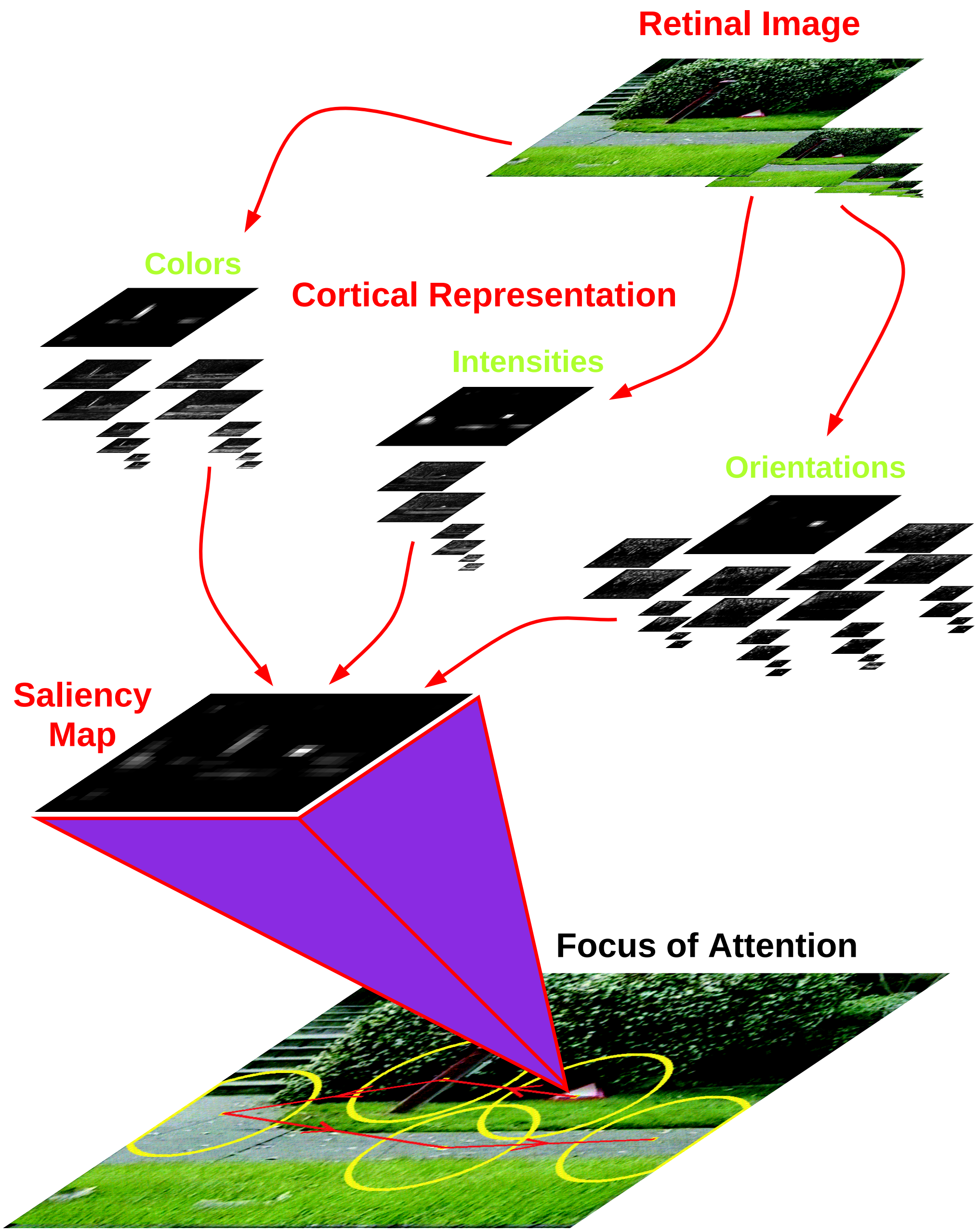


Figure 2.6: The multiscale feature maps, the conspicuity maps and the saliency map. In the intensity channel, six feature maps are computed, which simultaneously encode for on and off center-surround luminance contrast for the six center-surround scale pairs used (i.e., with center scale $c \in \{2, 3, 4\}$ and surround scale $s = c + \delta$, $\delta \in \{3, 4\}$). Six maps in the red/green color channel simultaneously encode for red/green and green/red double-opponency. Similarly, six maps in the red/green color channel simultaneously encode for red/green and green/red double-opponency. Finally, six maps for each of four orientations $(0, 45, 90$ and $135^\circ)$ detect local orientation contrast and edges. Shown at a larger magnification with the feature maps is a "conspicuity map" for each feature type, which combines information across all feature maps for that feature type. How the 42 feature maps can be combined into the single scalar saliency map is the topic of the next Section.

into a unique scalar measure of salience. One major difficulty in such combination resides in the fact that the different feature maps arise from different visual modalities, which encode for *a priori* not comparable stimulus dimensions: For example, how should a 10° orientation discontinuity compare to a 5% intensity contrast?

In addition, because of the large number of maps being combined, the system is faced with a severe signal-to-noise ratio problem: A salient object may only elicit a strong peak of activity in one or a few feature maps, tuned to the features of that object, while a larger number of feature maps, for example tuned to the features of distracting objects, may show strong peaks at numerous locations. For instance, a stimulus display containing one vertical bar among many horizontal bars yields an isolated peak of activity in the map tuned to vertical orientation at the scale of the bar; the same stimulus display, however, also elicits strong peaks of activity, in the intensity channel, at the locations of all bars, simply because each bar has high intensity contrast with the background. When all feature maps are combined into the saliency map, the isolated orientation pop-out hence is likely to be greatly weakened, at best, or even entirely lost, at worst, among the numerous strong intensity responses.

In what follows, we propose several feature combination strategies: Naive summation, learned linear combination, contents-based global non-linear amplification, and iterative localized interactions. Next, we will pursue our description of the model by assuming that one of these four schemes has been selected, and by generically denoting by the **"feature normalization operator $\mathcal{N}(.)$"** the computational processes applied to each feature map before it is fed to the inputs of the saliency map. In the next Chapter, we present a detailed comparison of the four versions of our model associated with these feature combination strategies.

2.3.1 Naive Summation

The most simple approach to solve this problem is to normalize all feature maps to the same total dynamic range (e.g., between 0 and 1), and to sum all feature maps into the saliency map. In this simple case, the feature normalization operator $\mathcal{N}(.)$ is the identity operator (and it is assumed that the feature detection mechanisms have been calibrated such as to yield similar response ranges). This strategy, which does not impose any *a priori* weight on any feature type, is referred to in what follows as the "Naive" strategy.

2.3.2 Learning Linear Combinations

Supervised learning can be introduced when specific targets are to be detected in the images presented to the model. In such case, each feature map is globally multiplied by a weighting factor. The final input to the saliency map is then the point-wise sum of all such feature maps. The feature normalization operator in this case consequently corresponds to a non-topographic multiplication by a scalar number. For each feature map, this number, the map's weight, is determined from a number of example images in which desired targets have been outlined. The training procedure then consists of increasing the weights of those maps which respond the targets better than to anything else, while decreasing the weights of those maps which either respond poorly to the targets, or respond better to non-target objects.

All feature map weights are trained simultaneously, based on a comparison, for each feature type, of the map's response inside and outside manually outlined image regions which contain the desired targets. A binary mask is hence first created for each image in the training set, in which 1 represents target regions and 0 non-target regions. In order to account for the fact that, at lower spatial scales, the boundaries of the targets in the feature maps may not exactly correspond to the high-resolution target outlines, the binary target masks are subsequently transformed into fuzzy representations: A 3/4 chamfer distance map $\mathcal{D}$ is created (Borgefors, 1991) from the target contours, i.e., a gray-level map encoding, at each image location, an approximation of the Euclidean distance between this location and the closest target region. Using this distance map, in what follows we will consider as being inside the target regions all locations x where $\mathcal{D}(x) \leq \mathcal{D}_{in}$, with $\mathcal{D}_{in}$ fixed; similarly, we will consider as being outside the target regions all locations x where $\mathcal{D}(x) \geq \mathcal{D}_{out}$, with $\mathcal{D}_{out}$ fixed. The distance thresholds $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$, when larger than 0, allow locations directly at the boundaries of the target regions, and yielding unreliable responses at coarse scales, to be ignored by the training procedure **(Fig. 2.7)**.

The learning procedure for the weight $w(\mathcal{M})$ of a feature map $\mathcal{M}$ consists of the following:

1. Compute the global maximum M_{glob} and minimum m_{glob} of the map $\mathcal{M}$;
2. Compute its maximum M_{in} inside the manually outlined target region(s) and its maximum M_{out} outside the target region(s);
3. Update the weight following an additive learning rule independent of the map's dynamic range:

$$w(\mathcal{M}) \leftarrow w(\mathcal{M}) + \eta(M_{in} - M_{out})/(M_{glob} - m_{glob}) \tag{2.5}$$

   where $\eta > 0$ determines the learning speed. Only positive or zero weights are allowed.

This learning procedure promotes, through an increase in weights, the participation to the saliency map of those feature maps which show higher peak activity inside the target regions than outside; after training, only such maps remain in the system while others,

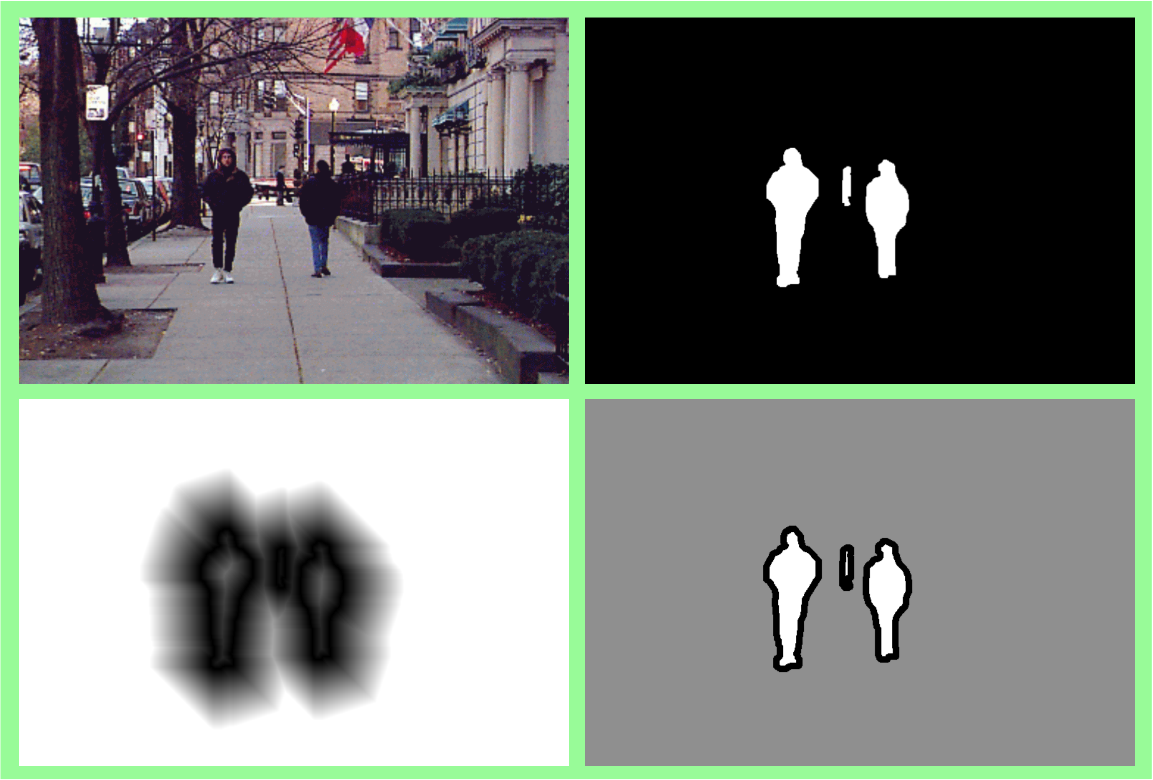


Figure 2.7: Selection of targets for supervised training. **Top-left:** Original image. **Top-right:** In this example, we wish to specialize the model for the task of detecting pedestrians. In this training image, the pedestrians have been manually outlined, and a binary target mask was created. **Bottom-left:** A 3/4 distance map computed from the contours of the target masks. The grey-level assigned to each pixel in this map represents the approximate Euclidean distance between that pixel and the nearest target contour. Note how this distance saturates (white color) far from the contours due to the limited grey-level range in the map. **Bottom-right:** Distance thresholds $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$ of about 3 pixels each have been applied to the distance map. The "inside" regions are shown in white, the "outside" region in grey, and the border regions which are ignored by the training algorithm in black.

whose weights converged to zero, are no more computed. The initial saliency map (before any attentional shift) is then scaled to a fixed range, such that only the relative weights of the feature maps are important; potential divergence of the additive learning rule (explosion of weights) is hence avoided by constraining the weights to a fixed sum.

Note that here we only consider local maxima of activity over various image areas, rather than the average activity over these areas. This is because local "peak" activity is what is important for visual salience: If a rather extended region contains only a very small but very strong peak of activity, this peak is highly salient and immediately "pops-out," while the average activity over the extended region may be low. This feature combination strategy is referred to in what follows as the "Trained" strategy.

2.3.3 Contents-Based Global Non-Linear Amplification

When no top-down supervision is available, we propose a simple normalization scheme, consisting of globally promoting those feature maps in which a small number of strong peaks of activity ("odd man out") are present, while globally suppressing feature maps eliciting comparable peak responses at numerous locations over the visual scene. The feature normalization operator $\mathcal{N}(.)$ then consists of the following **(Fig. 2.8)**:

1. Normalize all the feature maps to the same dynamic range, in order to eliminate across-modality amplitude differences due to dissimilar feature extraction mechanisms;

2. For each map, find its global maximum M and the average $\overline{m}$ of all the other local maxima;

3. Globally multiply the map by:

$$(M - \overline{m})^2. \tag{2.6}$$

Only local maxima of activity are considered such that $\mathcal{N}(.)$ compares responses associated with meaningful "activation spots" in the map and ignores homogeneous areas. Comparing the maximum activity in the entire map to the average over all activation spots measures how different the most active location is from the average. When this difference is large, the most active location stands out, and we strongly promote the map. When the difference is small, the map contains nothing unique and is suppressed. This contents-based non-linear normalization coarsely replicates the biological mechanism of localized excitation and lateral inhibition, in which neighboring similar features inhibit each other (Cannon & Fullenkamp, 1996). This feature combination strategy is referred to in what follows as the "Global non-linear amplification" strategy.

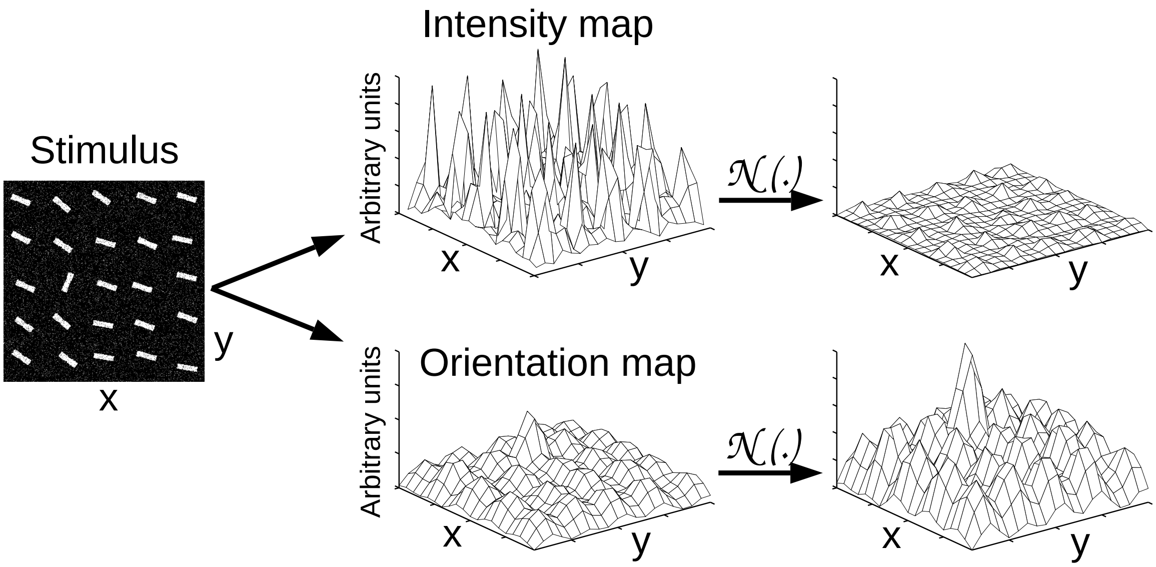


Figure 2.8: Contents-based global non-linear amplification. This operator determines a non-topographic feature map weight from the difference between the strength of the strongest peak of activity and the average strength of other peaks of activity in the map. Consequently, maps in which one or a few locations stand out will be amplified, while maps in which many locations show similar activity will be suppressed. Such operator is one of the key elements of our model, because it allows it with the severe signal-to-noise ratio faced when attempting to combine 42 feature maps into a single scalar saliency map. For example, in the image shown in this figure, intensity maps will show numerous strong responses, because each white bar has maximum contrast to the black background. Such strong activity can in many cases mask the relatively weak orientation pop-out seen in the oriented map, unless it is suppressed by an operator such as the one proposed here.

2.3.4 Iterative Localized Interactions

The global non-linear normalization presented in the previous Section is computationally very simple and is non-iterative, which easily allows for real-time implementation. However, it suffers from several drawbacks. First, this strategy is not very biologically plausible, since global computations, such as finding the global maximum in the image, are used, while it is known that cortical neurons are only locally connected. Second, this strategy has a strong bias towards enhancing those feature maps in which a unique location is significantly more conspicuous than all others. Ideally, each feature map should be able to represent a sparse distribution of a few conspicuous locations over the entire visual field; for example, our global non-linear amplification would suppress a map with two equally strong spots and otherwise no activity, while a human would typically report that both spots are salient.

Finally, the computational strategy employed in the previous Section is not robust to noise, when noise can be stronger than the signal (e.g., speckle or "salt-and-pepper" noise); in such stimuli, a single pixel of noise so high that it is the global maximum of the map would determine the map's scaling (such problem is however unlikely, since feature maps usually are built, from the noisy input image, using feature extraction mechanisms optimized to filter out the noise).

In this Section, we derive a generic model which does not impose any strong bias for any particular feature dimension. To this end, we implemented a simple within-feature spatial competition scheme, directly inspired by physiological and psychological studies of long-range cortico-cortical connections in early visual areas. These connections, which can span up to 6-8mm in striate cortex, are thought to mediate "non-classical" response modulation by stimuli outside the cell's receptive field. In striate cortex, these connections are made by axonal arbors of excitatory (pyramidal) neurons in layers III and V (Rockland & Lund, 1983; Gilbert *et al.*, 1996; Gilbert & Wiesel, 1989; Gilbert & Wiesel, 1983). Non-classical interactions are thought to result from a complex balance of excitation and inhibition between neighboring neurons as shown by electrophysiology (Sillito *et al.*, 1995; Sillito & Jones, 1996; Levitt & Lund, 1997), optical imaging (Weliky *et al.*, 1995), and human psychophysics (Polat & Sagi, 1994a; Polat & Sagi, 1994b; Zenger & Sagi, 1996).

Although much experimental work is being deployed in the characterization of these interactions, a precise quantitative understanding of such interactions still is in the early stages (Zenger & Sagi, 1996). Rather than attempting to propose a detailed quantitative account of such interactions, our model hence simply reproduces three widely observed features of those interactions: First, interactions between a center location and its non-classical surround appear to be dominated by an inhibitory component from the surround to the center (Cannon & Fullenkamp, 1991), although this effect is dependent on the relative contrast between center and surround (Levitt & Lund, 1997). Hence our model focuses on non-classical surround inhibition. Second, inhibition from non-classical surround locations is strongest from neurons which are tuned to the same stimulus properties as the center (Sillito *et al.*, 1995; Knierim & van Essen, 1992; Ts'o *et al.*, 1986; Gilbert & Wiesel, 1989; Malach *et al.*, 1993; Malach, 1994). As a consequence, our model implements interactions within each individual feature map rather than between maps. Third, inhibition appears strongest at a particular distance from the center (Zenger & Sagi, 1996), and weakens both with shorter and longer distances. These three remarks suggest that the structure of non-classical interactions can be coarsely modeled by a two-dimensional difference-of-Gaussians (DoG) connection pattern **(Fig. 2.9)**.

The specific implementation of these interactions in our model is as follows: Each feature map is first normalized to a fixed dynamic range (between 0 and 1), in order to eliminate feature-dependent amplitude differences due to different feature extraction mechanisms. Each feature map is then iteratively convolved by a large 2-D DoG filter, the original image is added to the result, and negative results are set to zero after each iteration. The DoG filter, a section of which is shown in **Fig. 2.9**, yields strong local excitation at each visual location, which is counteracted by broad inhibition from neighboring locations. Specifically, we have:

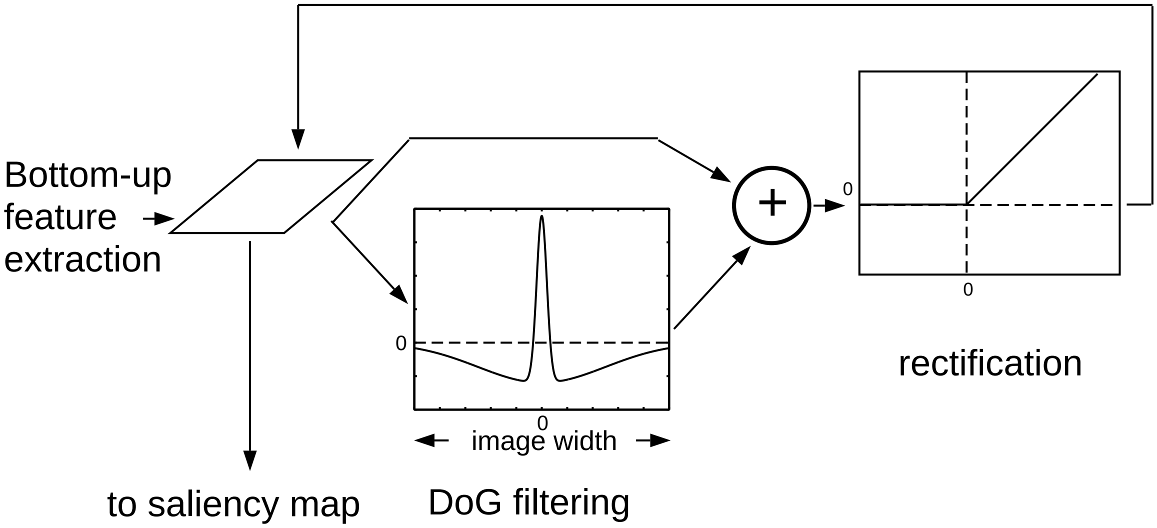
$$\mathcal{DoG}(x,y) = \frac{c_{ex}^2}{2\pi\sigma_{ex}^2} e^{-\frac{x^2+y^2}{2\sigma_{ex}^2}} - \frac{c_{inh}^2}{2\pi\sigma_{inh}^2} e^{-\frac{x^2+y^2}{2\sigma_{inh}^2}} \tag{2.7}$$

In our implementation, $\sigma_{ex} = 2\%$ and $\sigma_{inh} = 25\%$ of the input image width, $c_{ex} = 0.5$ and $c_{inh} = 1.5$ **(Fig. 2.9)**. At each iteration of the normalization process, a given feature map $\mathcal{M}$ is then subjected to the following transformation:

$$\mathcal{M} \leftarrow |\mathcal{M} + \mathcal{M} * \mathcal{DoG} - C_{inh}|_{\geq 0} \tag{2.8}$$

where $\mathcal{DoG}$ is the 2D Difference of Gaussian filter described above, $|.|_{\geq 0}$ discards negative values, and C_{inh} is a constant inhibitory term ($C_{inh} = 0.02$ in our implementation with the map initially scaled between 0 and 1). C_{inh} introduces a small bias towards slowly suppressing areas in which the excitation and inhibition balance almost exactly; such regions typically correspond to extended regions of uniform textures (depending on the DoG parameters), which we would not consider salient.

The 2D DoG filter, which is not separable, is implemented by taking the difference between the results of the convolution of $\mathcal{M}$ by the separable excitatory Gaussian of the DoG, and of the convolution of $\mathcal{M}$ by the separable inhibitory Gaussian. One reason for this approach is that two separable 2D convolutions (one of which, the excitatory Gaussian, has a very small kernel) and one subtraction are computationally much more efficient than one inseparable 2D convolution. A second reason is boundary conditions; this is an important



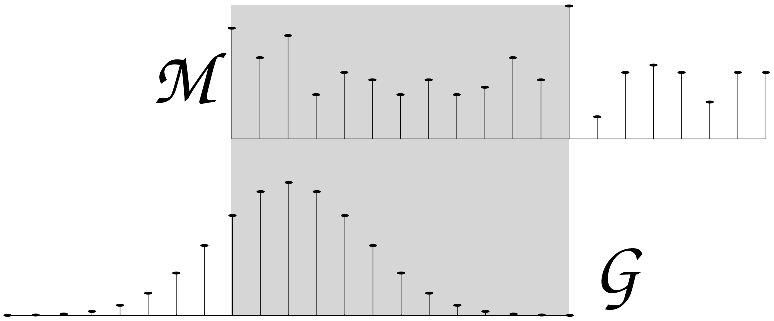


Figure 2.9: Iterative within-feature competition for salience. **(Top)** Illustration of the spatial competition for salience implemented within each of the 42 feature maps. Each map receives input from the linear filtering and center-surround stages. At each step of the process, the convolution of the map by a large Difference-of-Gaussians (DoG) kernel is added to the current contents of the map. This additional input coarsely models short-range excitatory processes and long-range inhibitory interactions between neighboring visual locations. The map is half-wave rectified, such that negative values are eliminated, hence making the iterative process non-linear. Ten iterations of the process are carried out before the output of each feature mapped is used in building the saliency map. **(Bottom)** “Truncated filter” boundary condition consists of only computing the dot product between filter $\mathcal{G}$ and map $\mathcal{M}$ where they overlap (shaded area), and of normalizing the result by the total area of $\mathcal{G}$ divided by its area in the overlap region.

problem here since the inhibitory lobe of the DoG is slightly larger than the entire visual field. Using Dirichlet (wrap-around) or "zero-padding" boundary conditions yields very strong edge effects which introduce unwanted non-uniform behavior of the normalization process (e.g., when using zero-padding, the corners of an image containing uniform random noise invariably become the most active locations, since they receive the least inhibition). We circumvent this problem by truncating the separable Gaussian filter $\mathcal{G}$, at each point during the convolution, to its portion which overlaps the input map $\mathcal{M}$ **(Fig. 2.9)**. The truncated convolution is then computed as, using the fact that $\mathcal{G}$ is symmetric around its origin:

$$\mathcal{M} * \mathcal{G}(x) = \frac{\sum_i \mathcal{G}(i)}{\sum_{i \in \{\text{overlap}\}} \mathcal{G}(i)} \sum_{i \in \{\text{overlap}\}} \mathcal{M}(i)\mathcal{G}(i) \qquad (2.9)$$

Using this "truncated filter" boundary condition yields uniform filtering over the entire image (see, e.g., **Fig. 2.10**), and, additionally, presents the advantage of being more biologically plausible than Dirichlet or zero-padding conditions: A visual neuron with its receptive field near the edge of our visual field indeed is not likely to implement zero-padding or wrap-around, but is likely to have a reduced set of inputs, and to accordingly adapt its output firing rate to a range similar to that of other neurons in the map.

Each feature map is subjected to 10 iterations of the process described in **Eq. 2.8**. The choice of the number of iterations is somewhat arbitrary: In the limit of an infinite number of iterations, any non-empty map will converge towards a single peak (except for a few unrealistic, singular configurations), hence constituting only a poor representation of the scene. With few iterations, however, spatial competition is weak and inefficient. Two examples of the time evolution of this process are shown in **Fig. 2.10**, and illustrate that using on the order of 10 iterations yields adequate distinction between the two example images shown. As expected, feature maps with initially numerous peaks of similar amplitude are suppressed by the interactions, while maps with one or a few initially stronger peaks become enhanced.

While the effect of this scheme seems similar to that of the previous scheme, the dynamics of this new scheme are however much more complex than those of $\mathcal{N}(.)$, since now the map is locally altered rather than globally (non-topographically) multiplied; for example, a map such as that at the top of **Fig. 2.10** converges to a single activated pixel (at the center of the initial strong peak) after a large number of iterations. Note finally that, although the range of the inhibitory filter seems to far exceed that of intrinsic cortico-cortical connections in primates (Gilbert & Wiesel, 1989), it is likely that such inhibition is fed back from higher cortical areas where receptive fields can cover substantial portions of the entire visual field, to lower visual areas with smaller receptive fields. In terms of implementation, the DoG filtering proposed here is best carried out within the multiscale framework of Gaussian

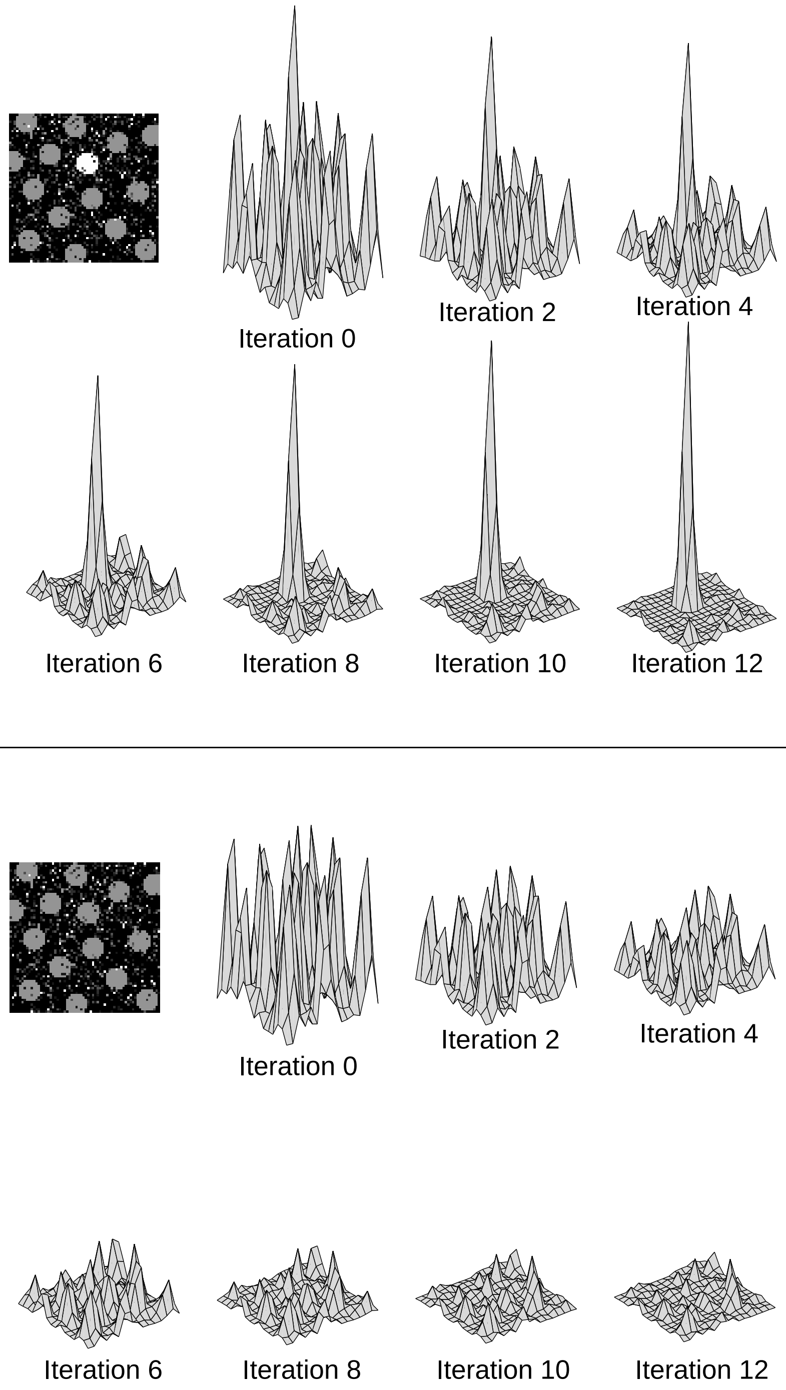


Figure 2.10: Demonstration of the iterative competition scheme. **(Top)** Iterative spatial competition for salience in a single feature map with one strongly activated location surrounded by several weaker ones. After a few iterations, the initial maximum has gained further strength while at the same time suppressing weaker activation regions. **(Bottom)** Iterative spatial competition for salience in a single feature map containing numerous strongly activated locations. All peaks inhibit each other more-or-less equally, resulting in the entire map being suppressed.

Pyramids (Itti *et al.*, 1998b). It is interesting to note that this within-feature spatial competition scheme resembles a "winner-take-all" network with localized inhibitory spread, which allows for a sparse distribution of winners across the visual scene (see (Horiuchi *et al.*, 1997) for a 1-D real-time implementation in Analog-VLSI).

2.4 The Conspicuity Maps

After normalization, the feature maps for intensity, color, and orientation are summed across scales into three separate "conspicuity maps," one for intensity, one for color and one for orientation **(Fig. 2.1b)**.

Feature maps are combined into three "conspicuity maps," $\overline{\mathcal{I}}$ for intensity (**Eq. 2.10**), $\overline{\mathcal{C}}$ for color (**Eq. 2.11**), and $\overline{\mathcal{O}}$ orientation (**Eq. 2.12**), at the scale ($\sigma = 4$) of the saliency map. They are obtained through across-scale addition, $\oplus$, which consists of reduction of each map to scale 4 and point-by-point addition:

$$\overline{\mathcal{I}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{I}(c,s)) \tag{2.10}$$

$$\overline{\mathcal{C}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [\mathcal{N}(\mathcal{RG}(c,s)) + \mathcal{N}(\mathcal{BY}(c,s))] \tag{2.11}$$

For orientation, four intermediary maps are first created by combination of the six feature maps for a given θ, and are then combined into a single orientation conspicuity map:

$$\overline{\mathcal{O}} = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} \mathcal{N}\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{O}(c,s,\theta)) \right) \tag{2.12}$$

The motivation for the creation of three separate channels, $\overline{\mathcal{I}}$, $\overline{\mathcal{C}}$ and $\overline{\mathcal{O}}$, and their individual normalization is the hypothesis that similar features compete strongly for saliency, while different modalities contribute independently to the saliency map. The three conspicuity maps are normalized and summed into the final input $\mathcal{S}$ to the saliency map:

$$\mathcal{S} = \frac{1}{3}\left(\mathcal{N}(\overline{\mathcal{I}}) + \mathcal{N}(\overline{\mathcal{C}}) + \mathcal{N}(\overline{\mathcal{O}})\right) \tag{2.13}$$

The motivation for the creation of three separate channels and their individual normalization is the hypothesis that similar features compete strongly for salience, while different modalities contribute independently to the saliency map. Although we are not aware of any supporting experimental evidence for this hypothesis, this additional step has the computational advantage of further enforcing that only a spatially sparse distribution of strong activity peaks is present within each visual feature type, before combination of all three types into the scalar saliency map.

2.5 The Saliency Map and Generation of Attentional Scanpaths

At any given time, the maximum of the saliency map corresponds to the most salient stimulus to which the focus of attention should be directed next, in order to allow for more detailed inspection by neurons along the occipito-temporal pathway. To find the most salient location, we have to determine the maximum of the saliency map.

This maximum is selected by application of a winner-take-all algorithm. Different mechanisms have been suggested for the implementation of neural winner-take-all networks (Koch & Ullman, 1985; Yuille & Grzywacz, 1989); in particular see (Tsotsos *et al.*, 1995) for a multi-scale version of the winner-take-all network. In our model, we used a two-dimensional layer of integrate-and-fire neurons with strong global inhibition in which the inhibitory population is reliably activated by any neuron in the layer (a more realistic implementation would consist of populations of neurons; for simplicity, we model such populations by a single neuron with very strong synapses). When the first of these integrate-and-fire cells fires (winner), it will generate a sequence of action potentials, causing the focus of attention (FOA) to shift to the winning location. These action potentials will also activate the inhibitory population, which in turn inhibits all cells in the layer, hence resetting the network to its initial state.

In the absence of any further control mechanism, the system described so far would direct its focus of attention, in the case of a static scene, constantly to one location, since the same winner would always be selected. To avoid this undesirable behavior, we follow Koch and Ullman (Koch & Ullman, 1985) and introduce inhibitory feedback from the winner-take-all (WTA) array to the saliency map. When a spike occurs in the WTA network, the integrators in the saliency map transiently receive additional input with the spatial structure of a difference of Gaussians. The inhibitory center (with a standard deviation of half the radius of the FOA) is at the location of the winner; it and its neighbors become inhibited in the saliency map. As a consequence, attention switches to the next-most conspicuous location **(Figs. 2.12, 2.13)**. Such an “inhibition of return” **(Fig. 2.11)** has been well demonstrated for covert attentional shifts in humans (Posner *et al.*, 1982; Kwak & Egeth, 1992). There is much less evidence for inhibition-of-return for eye movements in either humans or trained monkeys (Motter & Belky, 1998).

The function of the excitatory lobes (half width of four times the radius of the FOA) is to favor locality in the displacements of the focus of attention: If two locations are of nearly

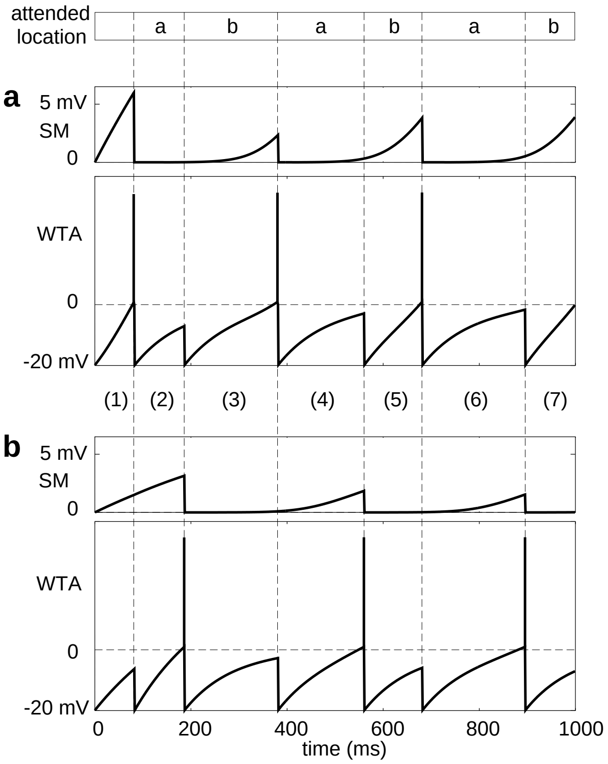


Figure 2.11: Dynamics of attentional focusing. Dynamical evolution of the potential of some simulated neurons in the saliency map (SM) and in the winner-take-all (WTA) networks. The input contains one salient location (a), and another input of half the saliency (b); the potentials of the corresponding neurons in the SM and WTA are shown as a function of time. During period **(1)**, the potential of both SM neurons (a) and (b) increases as a result of the input. The potential in the WTA neurons, which receive inputs from the corresponding SM neurons but have much faster time constants, increases faster. The WTA neurons evolve independently of each other as long as they are not firing. At about 80 ms, WTA neuron (a) reaches threshold and fires. A cascade of events follows: First, the focus of attention is shifted to (a); second, both WTA neurons are reset; third, inhibition of return (IOR) is triggered, and inhibits SM neuron (a) with a strength proportional to that neuron's potential (i.e., more salient locations receive more IOR, so that all attended locations will recover from IOR in approximately the same time). In period **(2)**, the potential of WTA neuron (a) rises at a much slower rate, because SM neuron (a) is strongly inhibited by IOR. WTA neuron (b) hence reaches threshold first. **(3)—(7):** In this example with only two active locations, the system alternatively attends to (a) and (b). Note how the IOR decays over time, allowing for each location to be attended several times. Also note how the amount of IOR is proportional to the SM potential when IOR is triggered (e.g., SM neuron (a) receives more IOR at the end of period (1) than at the end of period (3)). Finally, note how the SM neurons do not have an opportunity to reach threshold (at 20 mV) and to fire (their threshold is ignored in the model). Since our input images are noisy, we did not explicitly incorporate noise into the neurons' dynamics.

equal conspicuity, the one closest to the previous focus of attention will be attended next. This implementation detail directly follows the idea of "proximity preference" proposed by Koch and Ullman (Koch & Ullman, 1985).

The time constants, conductances, and firing thresholds of the simulated neurons are chosen so that the FOA jumps from one salient location to the next in approximately 30–70 ms (simulated time; (Saarinen & Julesz, 1991)), and so that an attended area is inhibited for approximately 500–900 ms (see **Fig. 2.12**). These delays vary for different locations with the strength of the saliency map input at those locations. The FOA therefore may eventually return to previously attended locations, as it is observed psychophysically. These simulated time scales are related to the dynamical model of integrate-and-fire neurons used in our model (see `http://www.klab.caltech.edu/~itti/` for the implementation source code, which clearly specifies all parameters of the simulated neurons using SI units).

2.6 Outlook

In this Chapter, we have proposed a conceptually simple model for the bottom-up control of visual attention in primates. Our model is based on a coarse replication of the early levels of biological visual processing. When possible, fast implementations were used for the various components of the model, which further makes it only an approximation of neurobiological vision systems.

The model implements a number of simple visual features and relies on two major computational principles: First, we use a single topographic saliency map to guide attention, and, second, a substantial amount of computation is done at the earliest stages of visual processing, in particular when our iterative feature normalization is used.

In the next Chapter, we will see that this simple model demonstrates surprisingly powerful performance at analyzing real color scenes. In the following Chapter we analyze some of the key components of the model which are responsible for this success and propose directions for further research.

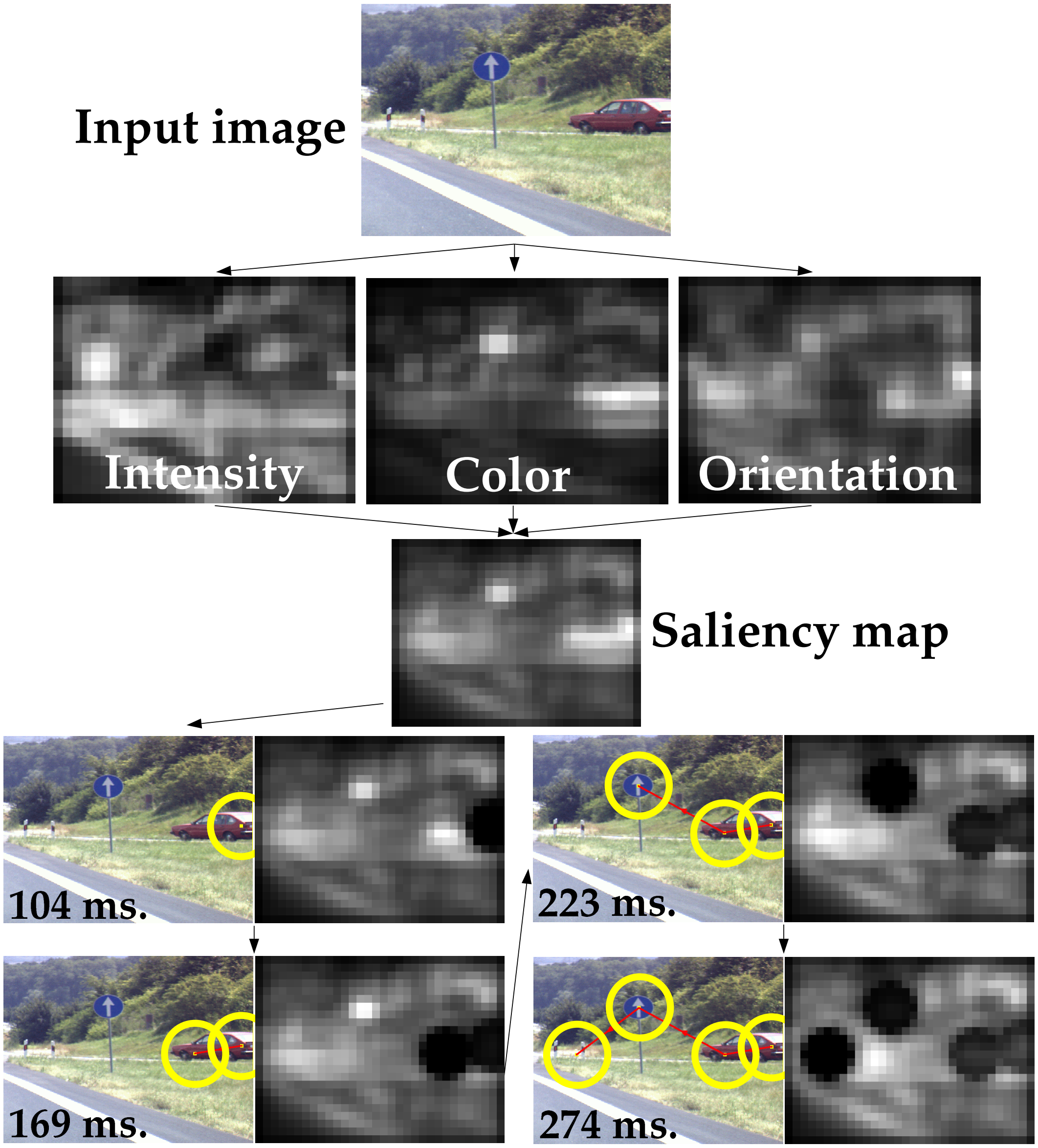


Figure 2.12: Demonstration of the model using global non-linear amplification. The input image had a resolution of 512 × 384 pixels in 24-bit color. Feature maps are extracted from the input image at several spatial scales, and are combined into three separate conspicuity maps (Intensity, Color and Orientation) at scale 4 (32 × 24 pixels). The three conspicuity maps that encode for saliency within these three domains are combined and fed into the single saliency map (also 32 × 24 pixels). A neural winner-take-all network then successively selects, in order of decreasing saliency, the attended locations. Once a location has been attended to for some brief interval, it is transiently suppressed in the saliency map by the inhibition of return mechanism (dark round areas). Note how the inhibited locations recover over time (e.g., the first attended location has regained some activity at 274 ms), due to the integrative properties of the saliency map. The radius of the focus of attention was 64 pixels.

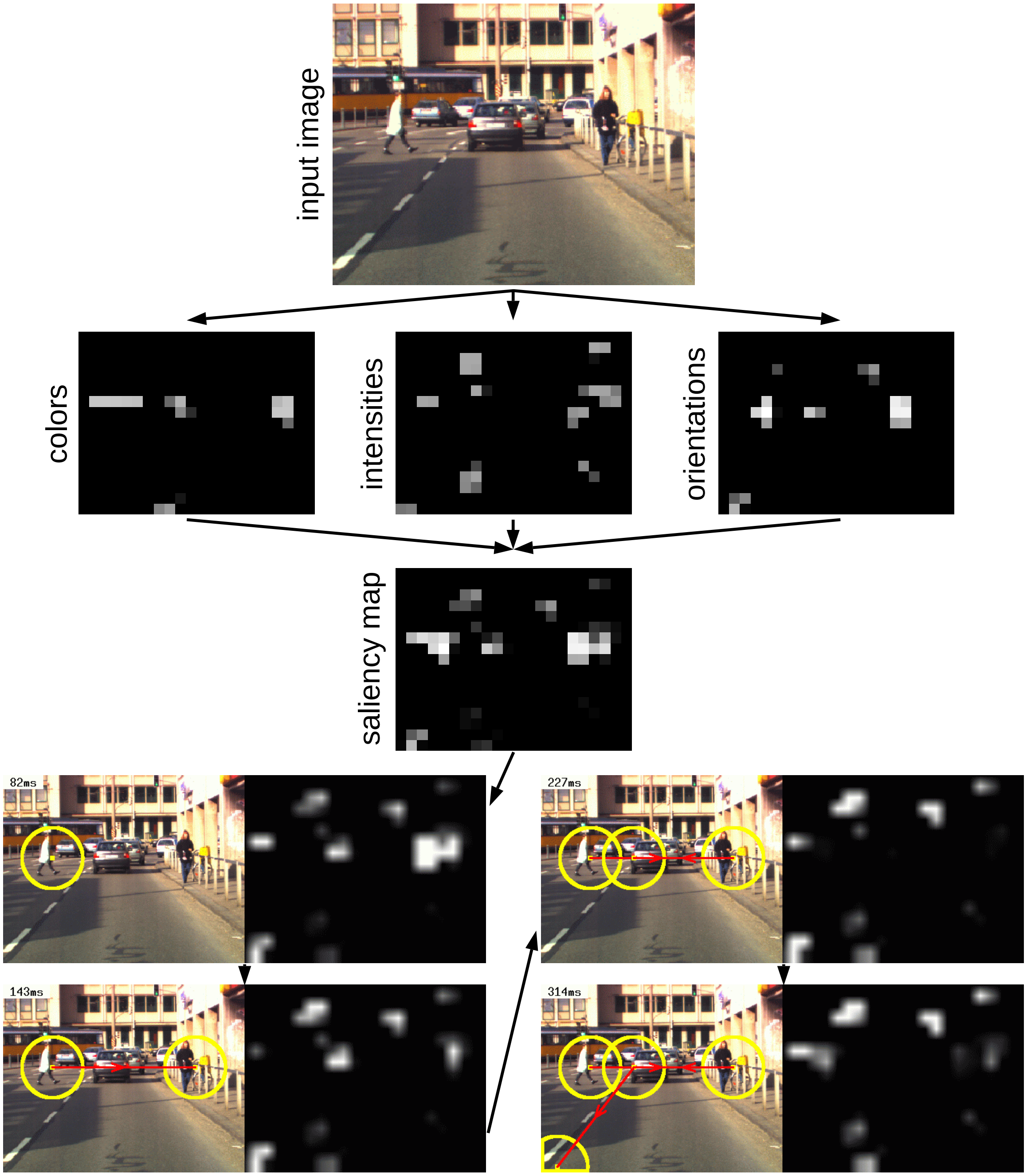


Figure 2.13: Demonstration of the model using iterative competition for salience. This figure follows the same principle as the previous one; one difference, however, is the feature normalization operator used, here the iterative competition for salience scheme. Note how both the feature maps and the saliency map are much sparser than with the global non-linear amplification scheme used in the previous figure. A systematic comparison of the four feature combination strategies proposed in this Chapter is detailed in the following Chapter.

Chapter 3 Applications of the Model

3.1 Overview

One of the most stringent tests for any artificial vision system is to evaluate its performance on real scenes. Such evaluation of our model is the main objective of the present Chapter.

We start by testing the model and its individual components with a number of simplified laboratory stimuli. We show that the model is capable of reproducing some aspects of human visual search when using simple arrays of prototypical visual patterns.

We then apply the model to a variety of digitized color images representing natural indoor and outdoor scenes. We use a wide variety of images, typically from databases submitted to us by external collaborators. This means that, rather than using very high quality images all acquired using a controlled and noise-free process, we evaluate our model on a realistic sample of images, ranging from low-resolution frames acquired by a consumer-grade video camera mounted on a moving vehicle to very high-resolution scans of 35mm slides.

We will see throughout this Chapter and further discuss in the next Chapter that one of the keys to the successful performance of our model at analyzing such images is our inclusion of non-linear computation at the earliest levels of processing, in the form of non-classical long-range suppression.

3.2 Test Images

We have extensively tested the model with our own or others' test images. These usually simplified images allowed us to examine in detail the functioning of the different components of the model.

One of the simplest possible tasks is the detection of bright spots on dark backgrounds, or dark spots in bright backgrounds. This task is solved reliably by our model, and the focus of attention immediately jumps to such stimuli. If there is more than one such stimulus, the system scans them one-by-one, in the order of decreasing contrast from the background. The same is true for stimuli that have a color or orientation different from that of the background. Progressively more complex images, involving color, orientation or combinations of these features can then be studied. Some examples are shown in **Fig. 3.1**. In general, the model behaves in a sensible manner when using either the global non-linear amplification or the iterative competition for salience as feature combination strategies. Many such simple examples, however, do not work when using either naive summation or hand-crafted feature weights. In particular, the latter two methods are not robust to the strong speckle noise added to some images in **Fig. 3.1**.

Of particular interest is the feature normalization operator. In our experience, this component is key to the performance of the model with natural scenes; indeed, this operator

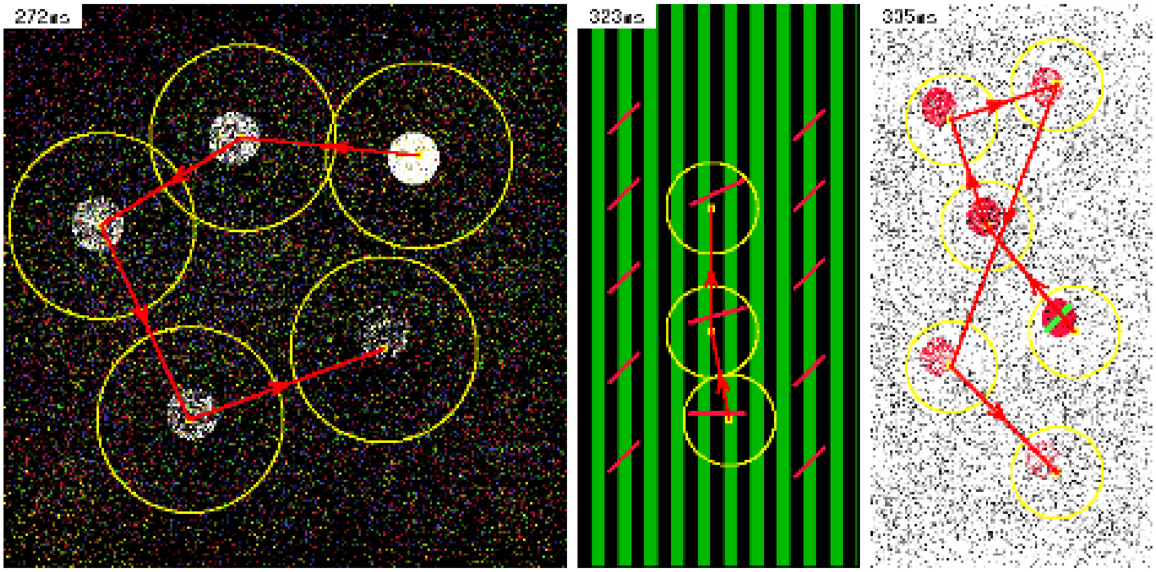


Figure 3.1: Model predictions for simple test images. When using either the global nonlinear amplification or the iterative competition for salience as feature combination strategies, attentional scanpaths generated by the model agree with our common-sense expectation; for example, of several oriented bars on a vertically oriented background, the most salient one should be that with largest orientation difference to the background. These simple results are however often not obtained when using either naive summation or handcrafted feature weights, especially when noise is present in the images.

allows us to discard all feature maps which contain noisy or unreliable responses over extended regions of the input image, and to enhance the feature maps which contain isolated salient locations **(Fig. 3.2)**.

The iterative competition for salience scheme yields best results, in particular when several distant locations are salient in the image to be analyzed. This operator is also of critical importance when high amounts of noise are present; in what follows, we show in a more systematic manner how an initially barely conspicuous target is either suppressed or made to stand out of its background through these interactions.

This simple experiment used only one feature map of 32x32 pixels, which contained one "target" and five "distractors." Both target and distractors were represented by disks of diameter 5 pixels. All stimulus images had background noise drawn from a uniform distribution of amplitude 25 (on an arbitrary scale also used below for target and distractor activities). One hundred stimuli were randomly generated for each condition. The "target to distractor ratio" (TDR) was defined as the activity at the target divided by the activity at the most salient distractor. In the first condition, initial distractor activity was 100 and initial target activity was 115 (initial TDR 1.15). For the 100 stimuli of that type, final TDR was 1.28 +/- 0.68 (mean +/- S.D.), i.e., the competitive process resulted in a strong increase of the TDR. In the second experiment, initial distractor activity was 200 and initial target activity was 215 (initial TDR 1.075), on the same arbitrary scale. For the 100 stimuli of that type, final TDR was 1.05 +/- 0.60, i.e., the competitive process resulted in a decrease in TDR. It is worth noting that the model predicted highly variable outcomes, as indicated by the S.D. of the final TDR; final TDR varied substantially from one instance to another, depending on the exact spatial configuration of target and distractors.

These experiments show that, although an initial TDR of 1.075 is insufficient for a target to pop-out, and yields global inhibition of all elements in the map including the target, an initial TDR of only 1.15 was sufficient to yield reliable enhancement of the target (to a final TDR of 1.28). These experiments hence demonstrate the double role of our iterative competition process, which suppresses regions of uniform activity and enhances locally conspicuous objects.

3.3 Pop-Out and Conjunctive Search

A first comparison of the model with humans can be made using the type of displays used in "Visual Search" tasks (Treisman, 1988). A typical experiment consists of a speeded alternative forced-choice task in which the presence of a certain item in the presented display has to be either confirmed or denied. It is known that stimuli which differ from nearby stimuli in a single feature dimension can be easily found in visual search, typically in a time which is nearly independent of the number of other items ("distractors") in the visual

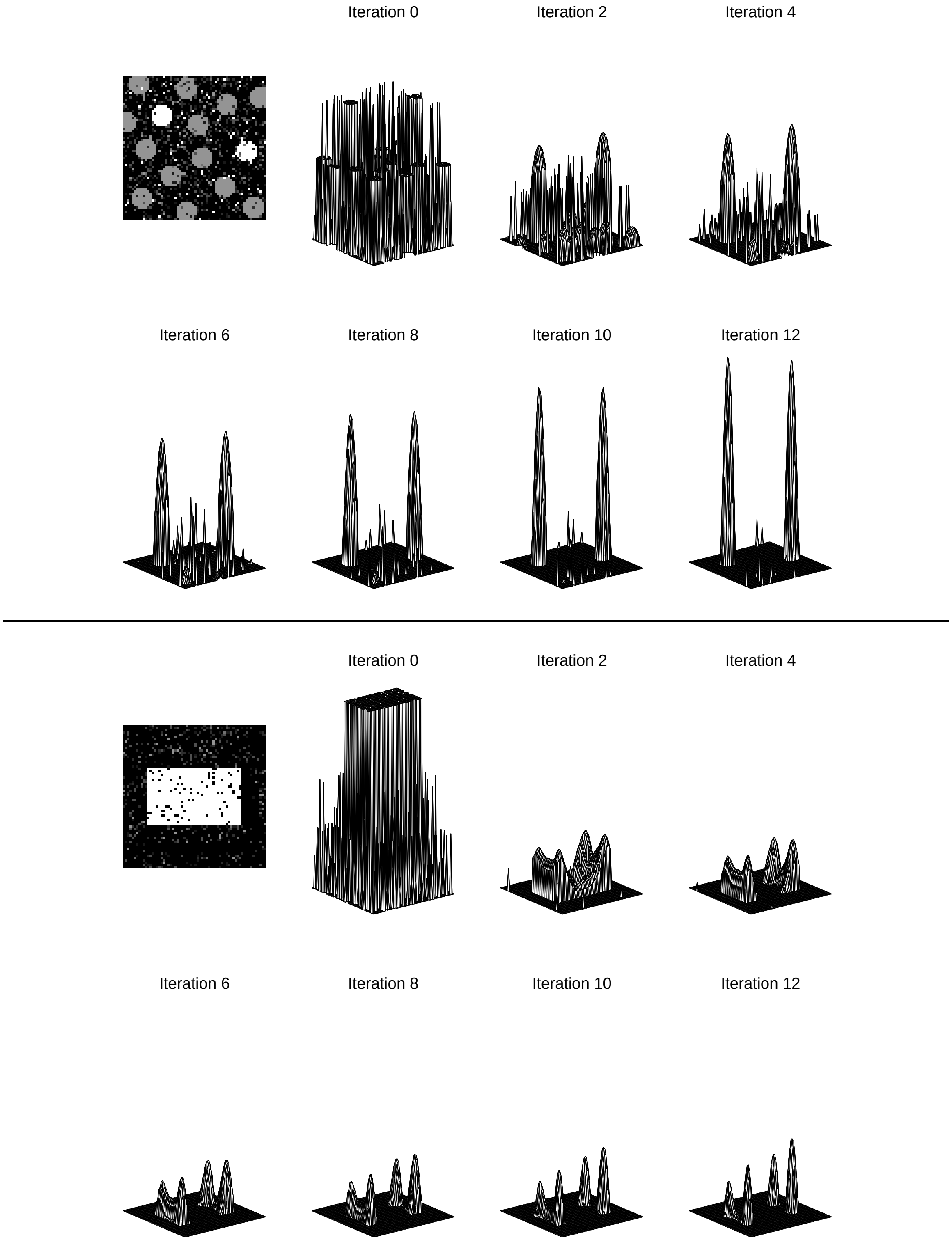


Figure 3.2: More examples of iterative competition for salience. Note how this operator allows several locations to simultaneously remain salient, while suppressing regions of uniform activity.

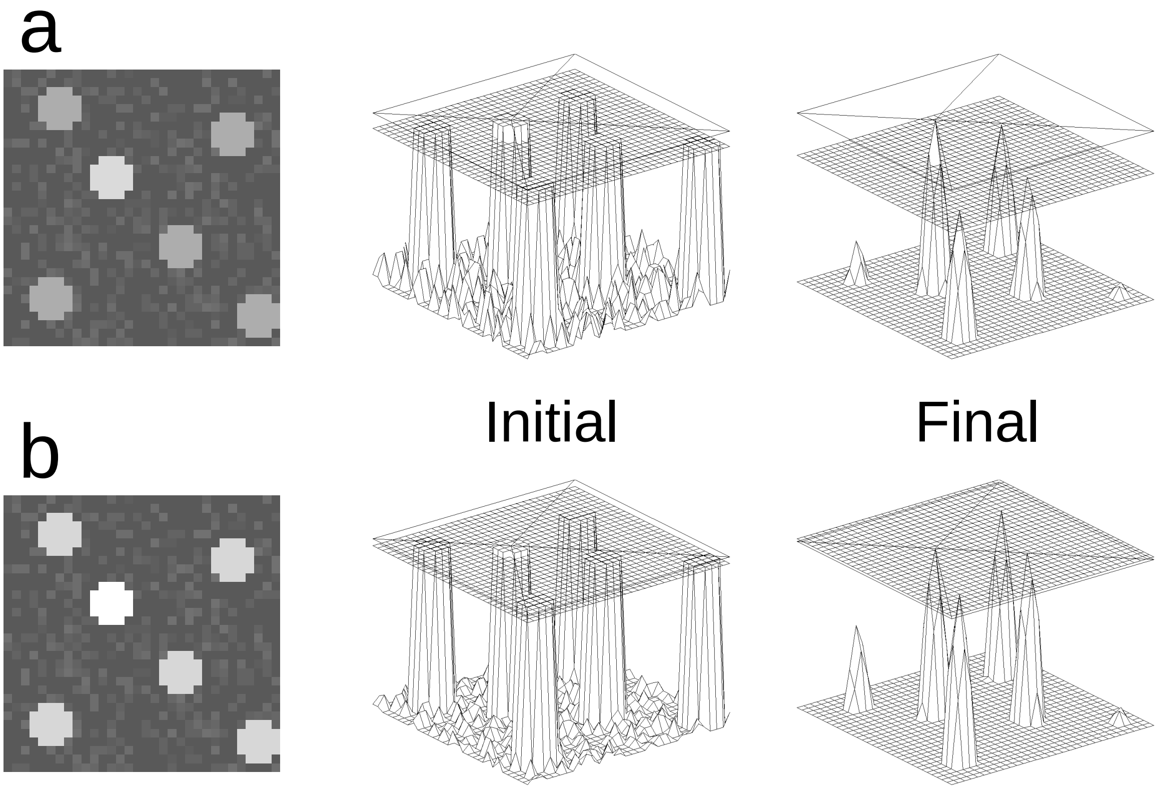


Figure 3.3: Illustration of how competition between salient locations in a feature map alters target-to-distractor saliency ratios. Gray-level images and middle-column plots represent the initial feature map, i.e., the presumed input from the preattentive feature extraction stages (grayscales have been distorted to make the difference between target and distractors visible). Right-column plots represent the feature map after spatial competition. All plots are scaled to the maximum activity in the map, in order to make relative differences obvious. In condition **(a)**, the target is initially 15% more active than the distractors; this relative difference is sufficiently large to allow the target to suppress all distractors more than the distractors can suppress the target. In this case, competition increases the initial difference between target and distractors. In condition **(b)**, the relative difference is now lower, at about 7.5%. Because of this low relative difference in saliency, target and distractors inhibit each other by approximately the same amount. As a result, the difference between target and distractors is attenuated further by the competitive process.

scene. In contrast, search times for targets which differ from distractors by a combination of features (a so-called "conjunctive task") are typically proportional to the number of distractors (Treisman & Gelade, 1980).

We generated three classes of synthetic images to simulate such experiments: (1) one red target (rectangular bar) among green distractors (also rectangular bars) with the same orientation; (2) one red target among red distractors with orthogonal orientation; and (3) one red target among green distractors with the same orientation, and red distractors with orthogonal orientation. In order not to artifactually favor any particular orientation, the orientation of the target was chosen randomly for every image generated. Also, in order not to obtain ceiling performance in the first two tasks, we added strong orientation noise to the stimuli (between -17 and +17 degrees with uniform probability) and strong color speckle noise to the entire image (each pixel in the image had a 15% uniform probability to become a maximally bright color among red, green, blue, cyan, purple, yellow and white). The positioning of the stimuli along a uniform grid was randomized (by up to ± 40% of the spacing between stimuli, in the horizontal and vertical directions), to eliminate any possible influence of our discrete image representations (pixels) on the system. Twenty images were computed for a total number of bars per image varying between 4 and 36, yielding the evaluation of a total of 540 images. In each case, the task of our model was to locate the target, whose coordinates were externally known from the image generation process, at which point the search was terminated. We are here not concerned with the actual object recognition problem within the focus of attention. The diameter of the FOA was fixed to slightly more than the longest dimension of the bars.

Results are presented in **Fig. 3.4** in terms of the number of false detections before the target was found. Clear pop-out was obtained for the first two tasks (color only and orientation only), independently of the number of distractors in the images. Slightly worse performance is found when the number of distractors is very small, which seems sensible since in these cases the distractors are nearly as salient as the target itself. Evaluation of these types of images without introducing any of the distracting noises described above yielded systematic pop-out (target found as the first attended location) in all images. The conjunctive search task showed that the number of shifts of the focus of attention prior to the detection of the target increased linearly with the number of distractors. Notice that the large error bars in our results indicate that our model usually finds the target either quickly (in most cases) or only after scanning a large number of locations.

3.4 Search Asymmetries

In some instances of visual search, where human observers detect the presence or absence of a special "target" visual pattern in an array of identical "distractor" patterns, search time

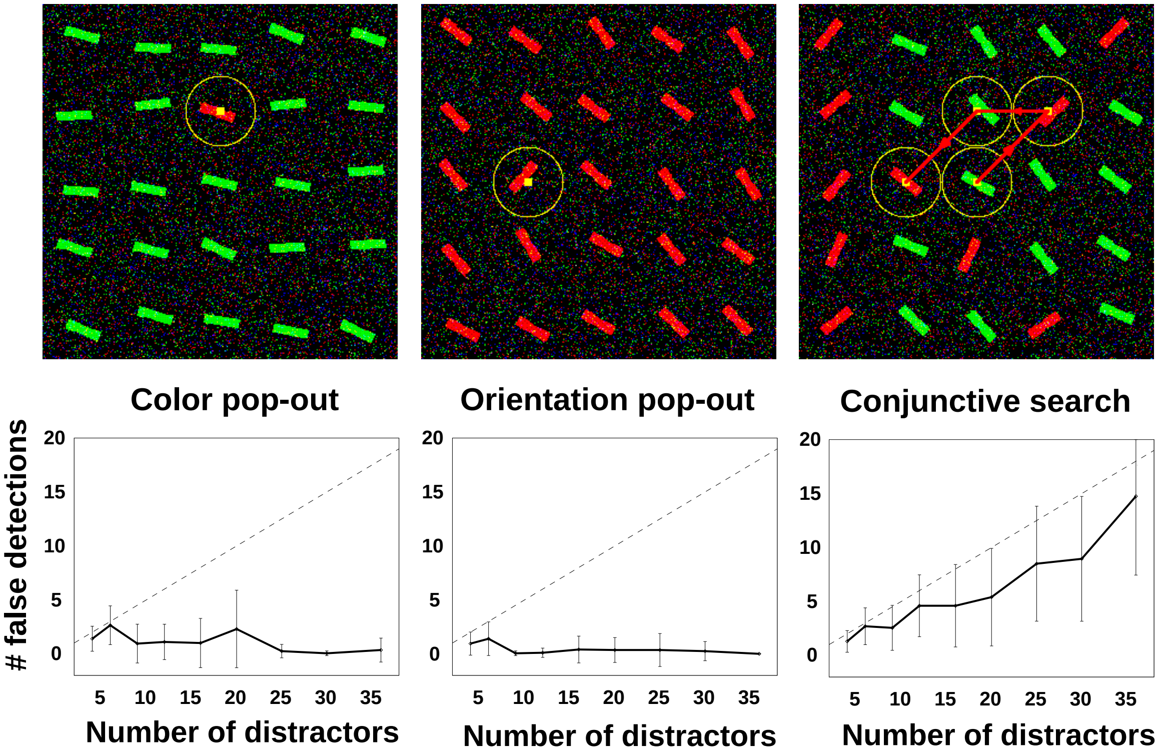


Figure 3.4: Pop-out and conjunctive search reproduced by our model. Model performance is shown on noisy versions of pop-out and conjunctive tasks of the type pioneered by Treisman and Gelade (1980). Stimuli consisted of arrays of isoluminant red and green colored bars with strong speckle noise added. Dashed lines: chance value, based on the size of the simulated visual field and the size of the candidate recognition area (corresponds to the performance of an ideal observer who scans, on average, half of the distractors prior to target detection). Solid lines: performance of the model. Error bars: one standard deviation. The typical search slopes of human observers in feature search and conjunction search, respectively, are successfully reproduced by the model. Each stimulus was drawn inside a 64 × 64 pixels box, and the radius of the focus of attention was fixed to 32 pixels. For a fixed number of stimuli, we tested 20 randomly generated images in each task; the saliency map and winner-take-all were initialized to zero (corresponding to a uniformly black visual input) prior to each trial.

asymmetries have been reported when target and distractor patterns are interchanged. A classical explanation for this finding is that targets which contain "richer features" are easier to find among simpler distractors than the opposite. For example, a curved line segment is detected faster among straight segments than the opposite, presumably because it has the added property of curvature. Our model seems ideally suited to testing this classical understanding of the origin of the asymmetries, because it allows us to directly look at the responses of the target and distractor patterns in individual feature maps. Consequently, we conducted simulations using search arrays with target and distractor elements which have been shown to yield search time asymmetries in humans **(Fig. 3.5)**.

The generation of stimuli for these simulations, and the analysis of the data, was performed by Farinaz Therani, who was a student with me during the summer of 1999.

Stimulus arrays were generated by an automatic program. Search elements were randomly jittered by up to 60% of their size and rotated by up to ±10°. Uniform color speckle noise with 10% probability was finally added. For 20 target/distractor pairs (e.g., "Q" among "O" or open among closed circles), we generated 20 instances of arrays containing 4x4 to 10x10 elements (seven sizes in total). The resulting 5600 images were evaluated by our model, and simulated search times were collected.

For control target/distractor pairs, for which no asymmetry is found in humans (e.g., red among green squares, or vertical among horizontal bars), the model did not either exhibit spurious asymmetries. For pairs yielding asymmetry in humans, the model generally reproduced the asymmetry. With some pairs, however, the model initially predicted an opposite asymmetry; careful examination of the model's internals revealed that such failure was due to luminance imbalance between target and distractor. After luminance correction, the correct asymmetries were obtained. In all asymmetry cases, the model showed significantly stronger activity in at least one feature map for the easily-found target **(Fig. 3.5)**. Our simulations hence confirm in a computational manner that asymmetries may be due to an "added property" in the target that is easy to detect.

It should be noted, however, that the asymmetries exhibited by the model often are much more extreme than what is observed in humans; for example, in the "Q" among "O" search of **Fig. 3.5**, "Q" is the first attended location when it is the target, while "O" is the last one when it is the target. This simply is due to the fact that our simple attention focusing mechanism, using a winner-take-all network to explore targets in order of decreasing salience, is not capable of detecting the presence of a locally empty space in the saliency map. Whether humans have this capacity or whether they simply find the non-salient target by chance is subject to further investigations.

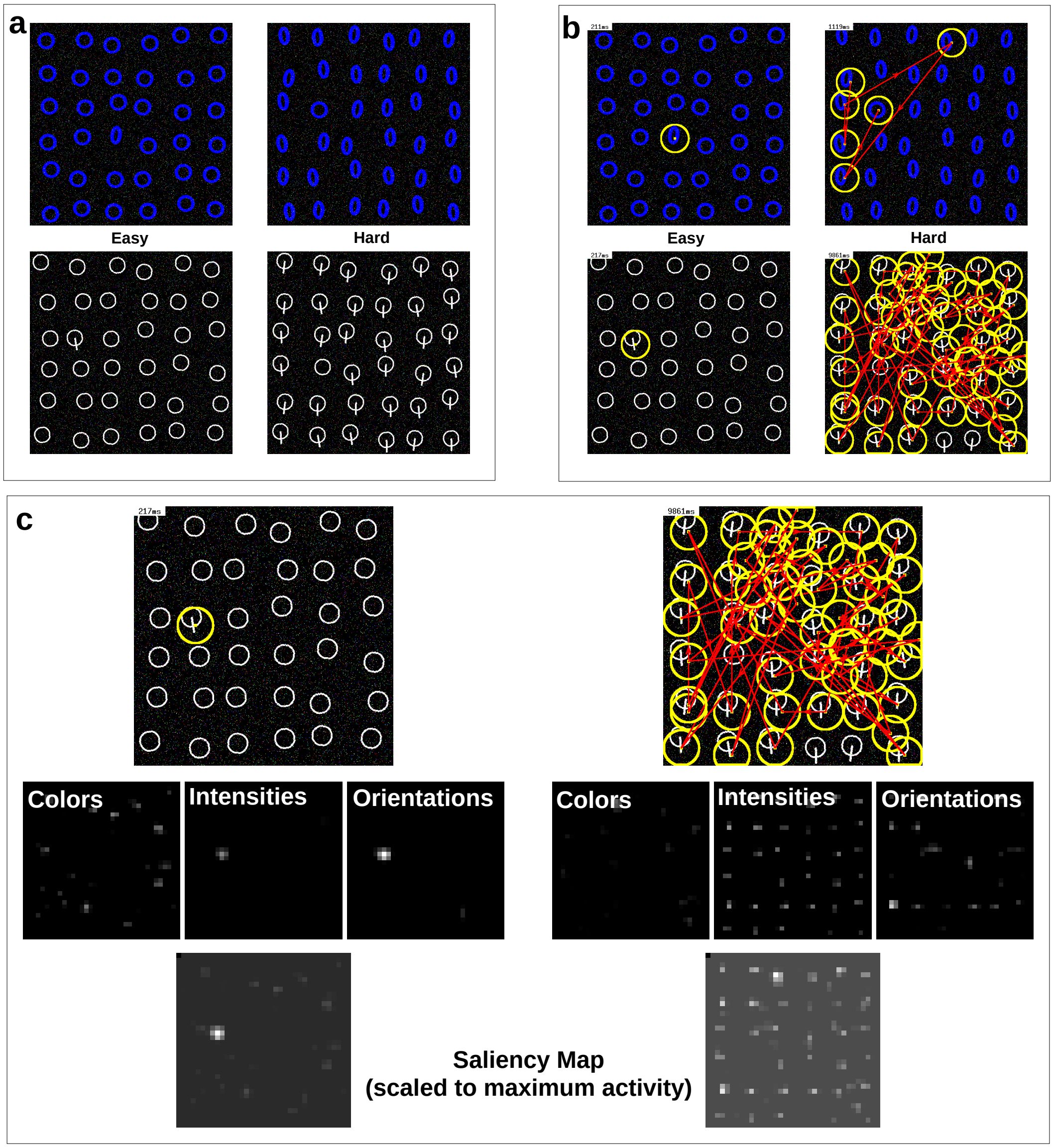


Figure 3.5: Search time asymmetries reproduced by our model. **(a)** Examples of search arrays used in our simulations. In humans, finding and ellipse among many circles is easier (as measured by faster reaction times) than finding a circle among many ellipses. Similarly, finding a letter “Q” among many letters “O” is easier than the opposite. **(b)** The model reproduces the basic asymmetry observed in humans. The model’s difference in search times is, however, extreme in some cases, while only a small difference in search time is usually observed in humans. **(c)** As expected, the search element which is easier to find (in this case, the letter “Q”) yields a much stronger activity when it is alone (i.e., when it is the target) in some feature maps (here, orientation, and to a lesser extent, intensity) than the search element which is harder to find. As a consequence, the “Q” among “O”s search yields a clear pop-out signal in the saliency map, while the “O” among “Q”s display yields very low salience for array elements (the saliency maps have been scaled to their maximum activity; the brighter background on the right map indicates lower overall activity).

3.5 Robustness to Image Noise

The model, using either the global non-linear amplification or the iterative competition for salience as feature combination strategies, proved very robust to the addition of noise to such images (**Fig. 3.6**). This was particularly true when the properties of the noise (e.g., its hue) were not directly conflicting with the main feature of the target.

Our model hence makes the prediction, which could be tested in human observers, that noise which is only weakly detected by the feature maps responsible for a target popping out should only minimally affect search time. If such result were to be verified in humans, it would argue, according to our model, for the fact that feature maps in different channels do not strongly interact. Indeed, it is because we have chosen in our model's architecture to separately treat color, intensity and orientation information in three non-interacting "conspicuity" maps that we find that black and white salt and pepper noise only minimally affects the detection time for a red target **(Fig. 3.6)**.

3.6 Search Performance in Complex Natural Scenes

We tested our model on a wide variety of real images, ranging from natural outdoor scenes to artistic paintings. All images were in color and contained significant amounts of noise, strong local variations in illumination, shadows and reflections, large numbers of "objects" often partially occluded, and strong textures. Most of these images can be interactively examined on the World-Wide-Web, at `http://www.klab.caltech.edu/~itti/attention/`. Overall, the results indicate that the system scans the image in an order which makes functional sense in most behavioral situations.

It should be noted however that it is not straightforward to establish objective criteria for the performance of the system with such images. Unfortunately, nearly all quantitative psychophysical data on attentional control are based on synthetic stimuli similar to those discussed in the next section. In addition, although the scan paths of overt attention (eye movements) have been extensively studied (Yarbus, 1967; Noton & Stark, 1971), it is unclear to what extent the precise trajectories followed by the attentional spotlight are similar to the motion of covert attention. Most probably, the requirements and limitations (e.g., spatial and temporal resolutions) of the two systems are related but not identical (Tsotsos *et al.*, 1995; Rao & Ballard, 1995). Although our model is mostly concerned with shifts of covert attention, and ignores all of the mechanistic details of eye movements, we attempt below a comparison between human and model target search times in complex natural scenes, using several databases of digitized images.

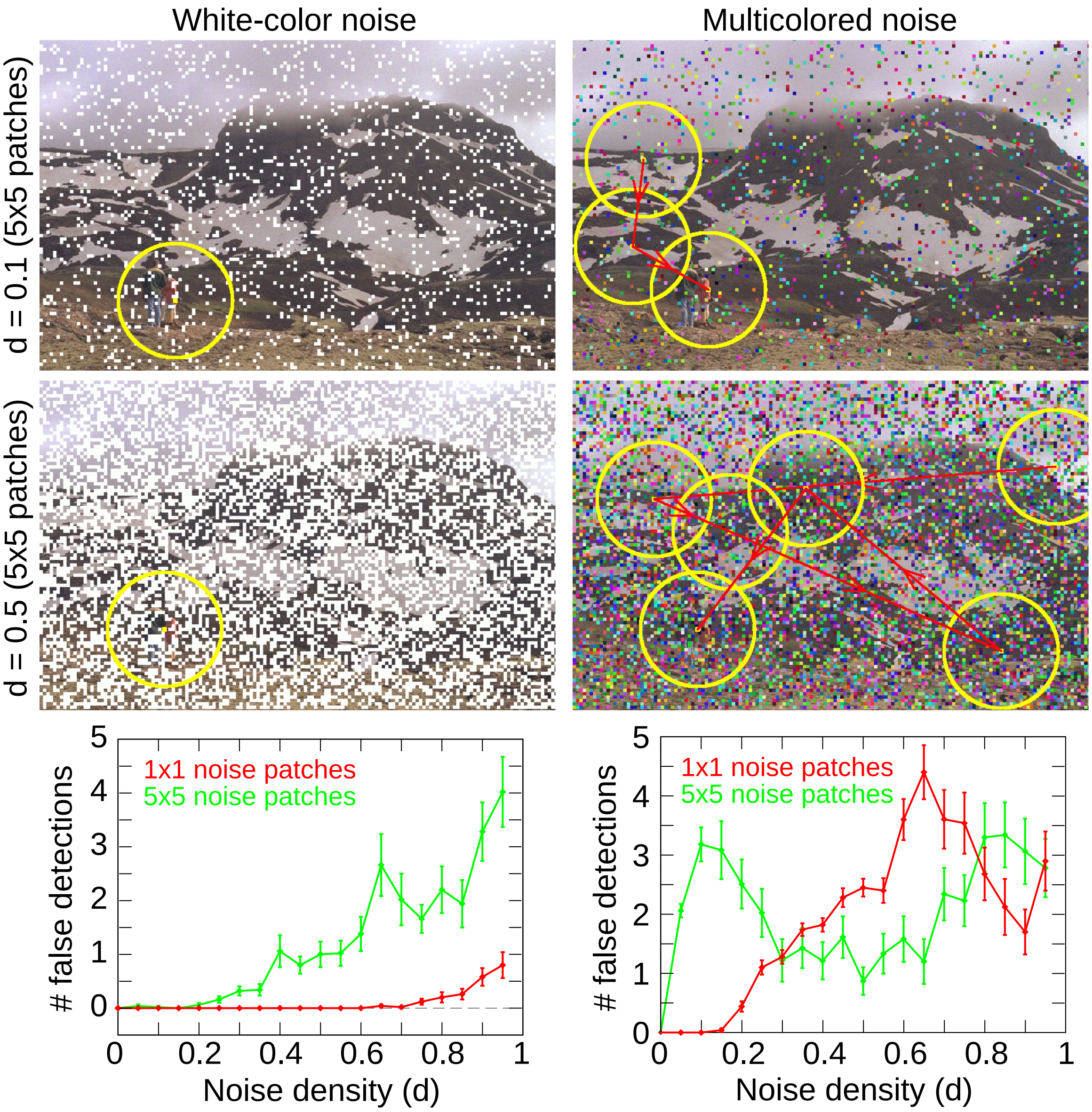


Figure 3.6: Influence of noise on detection performance, illustrated with a 768×512 scene in which a target (two people) is salient by its unique color contrast. The mean ± S.E. of false detections before target found is shown as a function of noise density for 50 instantiations of the noise. The system is very robust to noise which does not directly interfere with the main feature of the target (left; intensity noise and color target). When the noise has similar properties to the target, it impairs the target's saliency and the system first attends to objects salient for other features (here, coarse-scale variations of intensity).

3.6.1 Comparison of Feature Combination Strategies

We used three databases of natural color images to evaluate the different feature combination strategies proposed above **(Fig. 3.7)**. The first database consisted of images in which a red aluminum can is the target. It was used to demonstrate the simplest form of specialization, in which some feature maps in the system specifically encode for the main feature of the target (red color, which is explicitly detected by the system in a red/green feature map (Itti *et al.*, 1998b)). The second database consisted of images in which a vehicle's emergency triangle was the target. A more complicated form of specialization is hence demonstrated, since the target is unique in these images only by a conjunction of red color and of 0° (horizontal), 45° or 135° orientations. These four feature types are represented in the system by four separate and independent feature maps (Itti *et al.*, 1998b). The third database consisted of 90 images acquired by a video camera mounted on the passenger side of a vehicle driven on German roads, and contained one or more traffic signs. Among all 90 images, 39 contained one traffic sign, 35 contained two, 12 contained three, 2 contained four, and 1 contained five traffic signs.

All targets were outlined manually, and binary target masks were created. A target was considered detected when the focus of attention (FOA) intersected the target. The images were 640×480 (red can and triangle) and 512×384 (traffic signs) with 24-bit color, and the FOA was a disk of radius 80 (red can and triangle) and 64 (traffic signs) pixels. Complete coverage of an entire image would consequently require the FOA to be placed at 31 different locations (with overlap). A system performing at random would have to visit an average of 15.5 locations to find a unique, small target in the image.

Each image database was split into a training set (45 images for the can, 32 for the triangle, 45 for the traffic signs) and a test set (59, 32 and 45 images respectively). Learning consisted, for each training set, of 5 randomized passes through the whole set with halving of the learning speed η after each pass.

We compared the results obtained on the test image sets with the four proposed feature combination strategies:

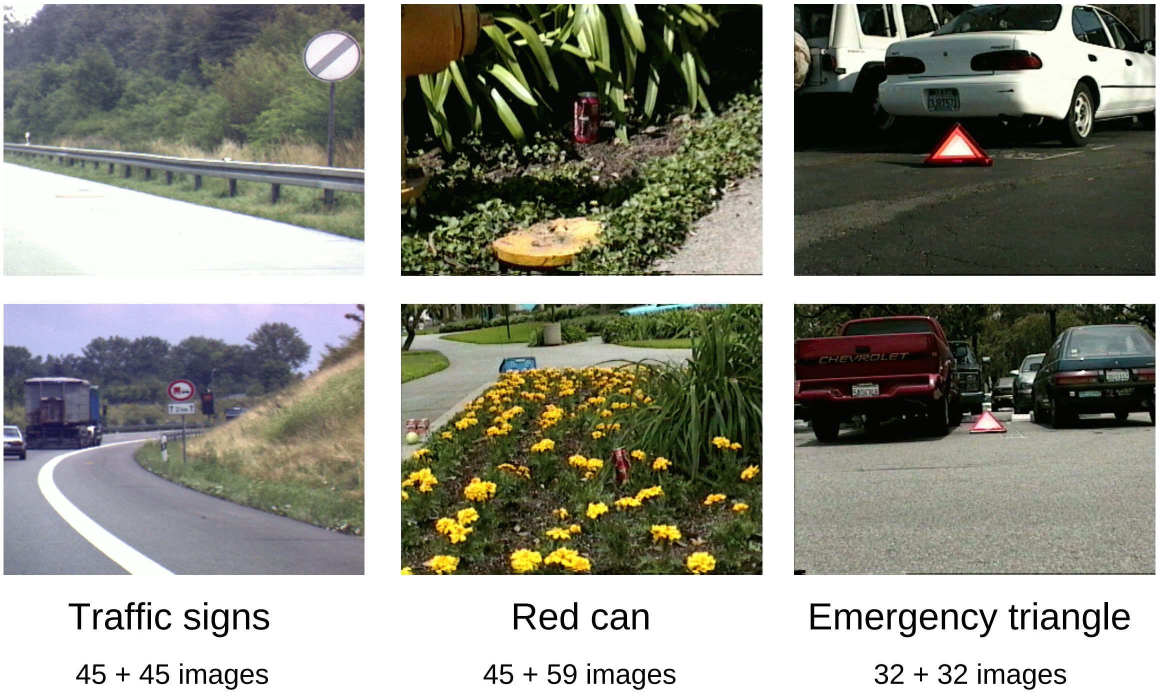
1. "Naive" model with no dedicated normalization and all feature weights set to unity;

2. Model with the "global non-linear amplification" normalization;

3. Model with 12 iterations of the "Iterative" normalization;

4. "Trained" model, i.e., with no dedicated normalization but feature weights learned from the corresponding training set.

We retained in the test sets only the most challenging images, for which the target was *not* immediately detected by at least one of the four versions of the model (easier



Traffic signs

45 + 45 images

Red can

45 + 59 images

Emergency triangle

32 + 32 images

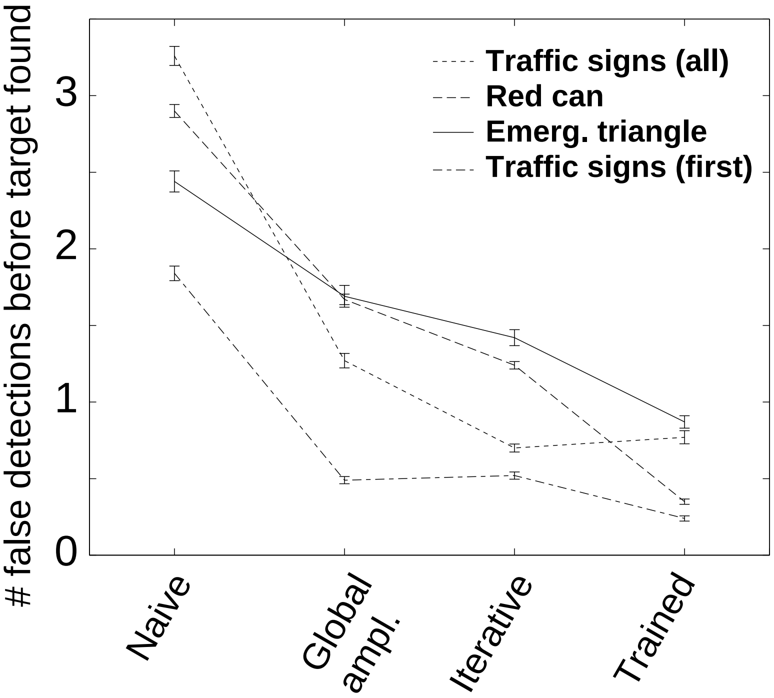
Figure 3.7: Example images from the three image databases studied. The number of images for training + test sets are shown for each database.

images in which at least one version of the model could immediately find the targets had been previously discarded to ensure that performance was not at ceiling). Results are summarized in **Table 3.1** and **Fig. 3.8**.

The naive model, which represents the simplest solution to the problem of combining several feature maps into a unique saliency map, performed always worse than when using global non-linear amplification. The simple contents-based non-linear normalization indeed proved particularly efficient at eliminating feature maps in which numerous peaks of activity were present, such as, for example, intensity maps in images containing large variations in illumination. Furthermore, the more detailed, iterative implementation of spatial competition for salience yielded comparable or better results, in addition to being more biologically plausible.

The additive learning rule also proved efficient in specializing the generic model. One should be aware, however, that only limited specialization can be obtained from such global weighting of the feature maps: Because such learning simply enhances the weight of some maps and suppresses others, poor generalization should be expected when trying to learn for a large variety of objects using a single set of weights, since each object would ideally require a specific set of weights. Additionally, the type of linear training employed here is limited, because *sums* of features are learned rather than *conjunctions*. For example, the model trained for the emergency triangle might attend to a strong oblique edge even if there was no red color present or to a red blob in the absence of any oblique orientation. To what extent humans can be trained to pre-attentively detect learned conjunctive features remains controversial (Niebur & Koch, 1998). Nevertheless, it was remarkable that the trained model performed best of the four models studied here for the database of traffic signs, despite the wide variety of shape (round, square, triangle, rectangle), color (red, white, blue, orange, yellow, green) and texture (uniform, striped, lettered) of those signs in the database.

In summary, while the “Naive” method consistently yielded poor performance and the “Trained” method yielded specialized models for each task, the iterative normalization operator, and its global non-iterative approximation, yielded reliable yet non-specific detection of salient image locations. We believe that the latter two represent the best approximations to human saliency among the four alternatives studied here. One of the key elements in the iterative method is the existence of a non-linearity (threshold) which suppresses negative values; as we can see in **Figs. 2.10** and **3.2**, in a first temporal period, the global activity over the entire map typically decreases as a result of the mutual inhibition between the many active locations, until the weakest activation peaks (typically due to noise) pass below threshold and are eliminated. Only after the distribution of activity peaks has become sparse enough can the self-excitatory term at each isolated peak overcome the inhibition received from its neighbors, and, in a second temporal period, the map’s global activity



	Naive	Global Ampl.	Iterative	Trained
Red can	2.90 ± 2.50	1.67 ± 2.01	1.24 ± 1.42	0.35 ± 1.03
Triangle	2.44 ± 2.20	1.69 ± 2.28	1.42 ± 1.67	0.87 ± 1.29
Traffic[1]	1.84 ± 2.13	0.49 ± 1.06	0.52 ± 1.05	0.24 ± 0.77
Traffic[2]	3.26 ± 2.80	1.27 ± 2.12	0.70 ± 1.18	0.77 ± 1.93

[1] *before* first *sign found.* [2] *before* all *signs found.*

Table 3.1: Average number of false detections (mean ± standard deviation) before target(s) found, for the Red Can test set (n=59), Emergency Triangle test set (n=32) and Traffic Signs test set (n=45; 17 images with 1 sign, 19 with 2, 6 with 3, 2 with 4 and 1 with 5). For the traffic sign images which could contain more than one target per image, we measured both the number of false detections before the first target hit, and before all targets in the image had been detected.

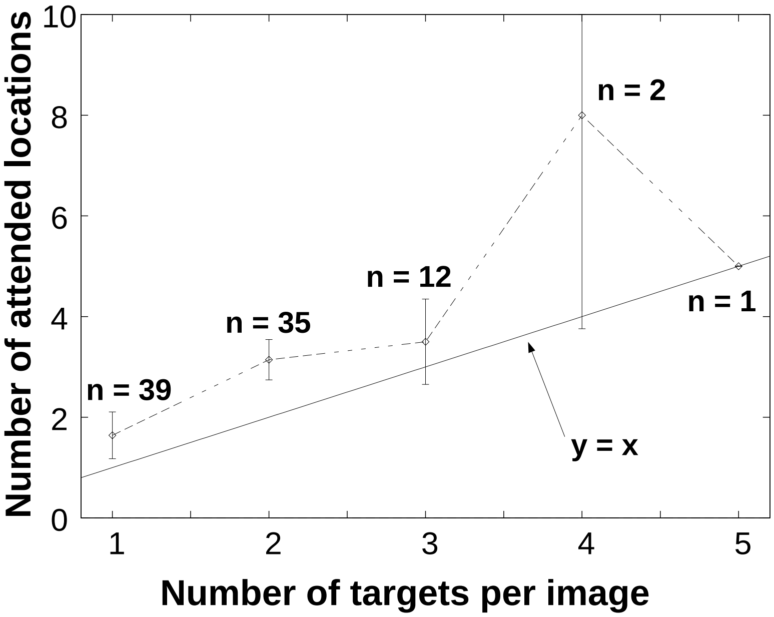


Figure 3.8: Breakdown of results for "Traffic Signs" database, according to the number of targets (signs) per image. Along the vertical axis is plotted the total number of attentional shifts generated by the model in order to find all targets. The straight line ($y = x$) represents performance of an optimal detector which finds one new target for each new shift. Our model performed very close to this optimal.

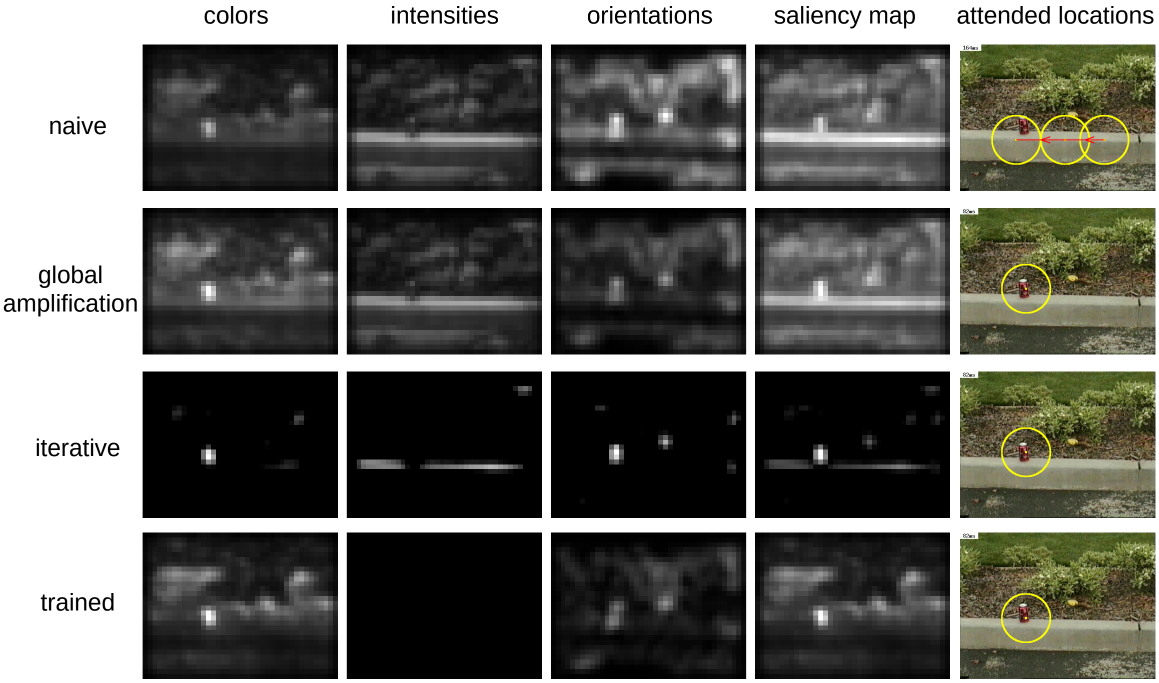


Figure 3.9: Comparison of the internals of the four versions of the model, for one image from the "red can" test set, in which a red aluminum can is the most salient object. The can appears with medium strength in the color maps, due to its color contrast with the background (the response is not the strongest possible because the background is not green, and only red/green and blue/yellow color contrasts are computed). The curb, however, appears very strongly in all intensity maps, and also less strongly in the horizontal orientation maps. In the naive version of the model, the color activity from the can is outnumbered by the activity elicited by the curb in a larger number of intensity and orientation maps. As a result, detection of the can is accidental, while the model is scanning the curb. The global non-linear amplification strategy yields strong suppression of the horizontal orientation, because more localized activation peaks exist in the vertical orientation, as well as some suppression of the extended curb in the intensity channel. The color channel, with its strong singularity, is however globally enhanced and yields correct detection of the can. The iterative strategy yields complete suppression of the horizontal orientation as well as overall much suppression of all regions which are not among the few strongest in each feature map. The red can clearly becomes the most salient location in the image. Finally, training using other images with similar views of this red target of vertical orientation has entirely suppressed the intensity and horizontal orientations, such that the saliency map is dominated by the color channel. The trained model hence easily finds the can as the most salient object.

starts increasing again. If many comparable peaks are present in the map, the first period of decreasing activity will be much slower than if one or a few much stronger peaks efficiently inhibits all other peaks. In **Fig. 3.9**, we show a comparison of the internal maps for the four versions of the model on a test image. This figure demonstrates, in particular, how the "Iterative" scheme yields much sparser maps, in which most of the noisy activity present in some channels (such as the intensity channel in the example image) is strongly suppressed.

3.6.2 Comparison to Spatial Frequency Content

By construction, the saliency map may be considered as topographically encoding for the probability for a generic observer to attend to given locations in the visual scene (considering only simple tasks and viewer-independent image features). It is hence important to quantitatively compare the model's attentional fixations to human attentional fixations, which unfortunately are not directly measurable. Although trajectories recorded with eye-tracking devices may differ substantially from attentional trajectories (Tsotsos *et al.*, 1995), it is reasonable to assume that the first few human eye fixations produced upon brief presentation of an image are mostly guided by bottom-up, saliency-driven mechanisms, before scene interpretation and other top-down mechanisms become active.

Reinagel and Zador (1999) used an eye-tracking device to analyze the local spatial frequency distributions along eye scan paths generated by humans while free-viewing grayscale images. They found the spatial frequency content at the fixated locations to be significantly higher than, on average, at random locations. Here we investigate whether our model would reproduce the findings of Reinagel and Zador. In the following two Sections, we present quantitative comparisons between human and model times to find a target in a complex natural scene, and preliminary experiments with eye movement recording done in our laboratory.

We constructed a simple measure of spatial frequency content (SFC): At a given image location, a 16 × 16 image patch is extracted from each $I(2), R(2), G(2), B(2)$ and $Y(2)$ map, and 2D Fast Fourier Transforms (FFTs) are applied to the patches. For each patch, a threshold is applied to compute the number of non-negligible FFT coefficients; the threshold corresponds to the FFT amplitude of a just perceivable grating (1% contrast). The SFC measure is the average of the numbers of non-negligible coefficients in the five corresponding patches. The size and scale of the patches were chosen such that the SFC measure is sensitive to approximately the same frequency and resolution ranges as our model; also, our SFC measure is computed in the RGB channels as well as in intensity, like the model. Using this measure, an SFC map is created at scale 4 for comparison with the saliency map (**Fig. 3.10**).

Model predictions were compared to the measure of local SFC, in an experiment similar

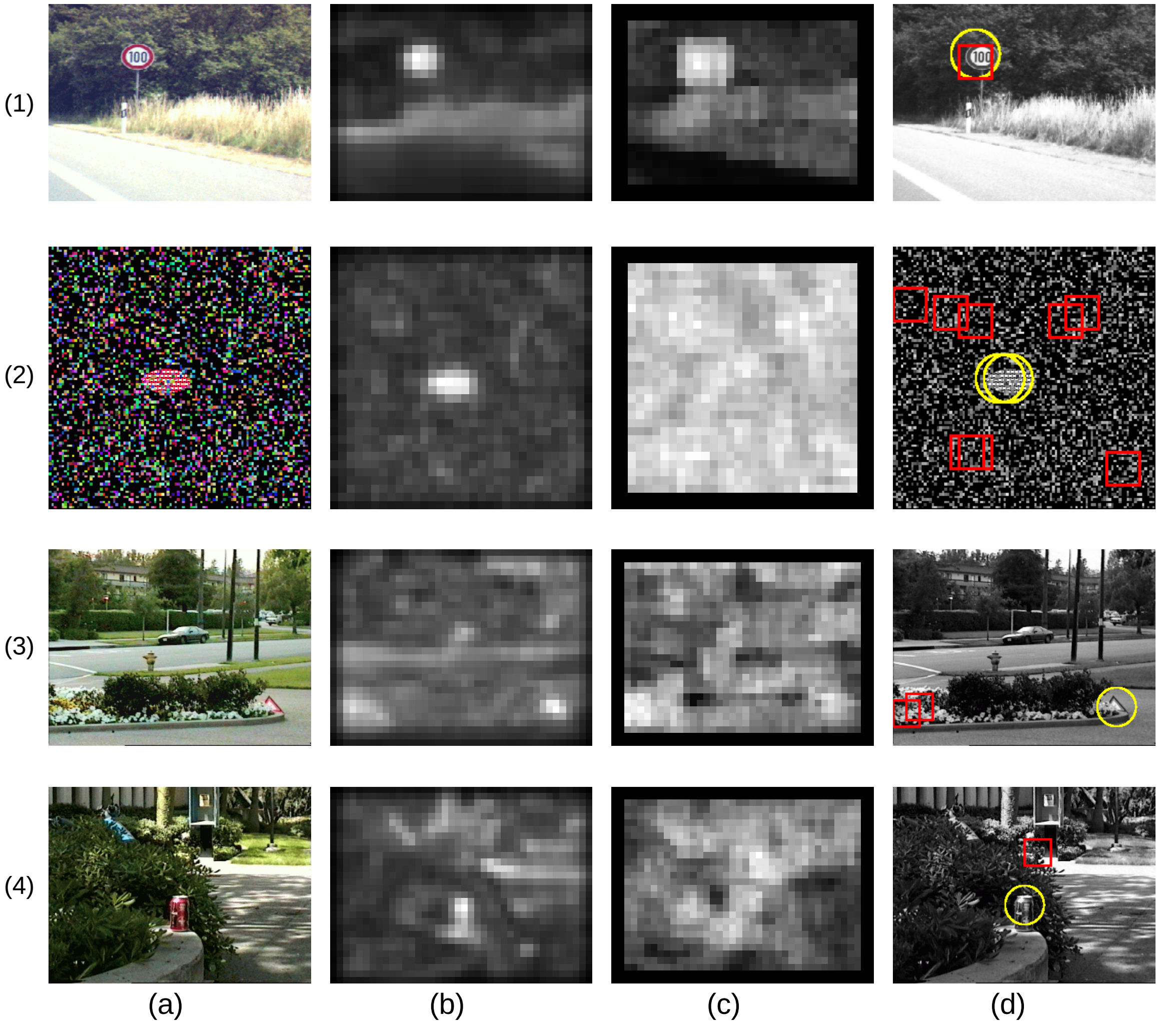


Figure 3.10: Comparison of the saliency map to the map of spatial frequency content. Examples of color images **(a)**, the corresponding saliency map inputs **(b)**, spatial frequency content (SFC) maps **(c)**, locations at which input to the saliency map was higher than 98% of its maximum **(d; yellow circles)**, and image patches for which the SFC was higher than 98% of its maximum **(d; red squares)**. The saliency maps are very robust to noise, while SFC is not.

to that of Reinagel and Zador (1999), using natural scenes with salient traffic signs (90 images), red soda can (104 images), or vehicle's emergency triangle (64 images). Similar to Reinagel and Zador's findings, the SFC at attended locations was significantly higher than the average SFC, by a factor decreasing from 2.5 ± 0.05 at the first attended location to 1.6 ± 0.05 at the 8th attended location. Although this result does not necessarily indicate similarity between human eye fixations and the model's attentional trajectories, it indicates that the model, like humans, is attracted to "informative" image locations, according to the common assumption that regions with richer spectral content are more informative. The SFC map was similar to the saliency map for most images (e.g., **Fig. 3.10.1**). However, both maps differed substantially for images with strong, extended variations of illumination or color (e.g., due to speckle noise): While such areas exhibited uniformly high SFC, they had low saliency because of their uniformity (**Figs. 3.10.2, 3.10.3**). In such images, the saliency map was usually in better agreement with our subjective perception of saliency. Quantitatively, for the 258 images studied here, the SFC at attended locations was significantly lower than the maximum SFC, by a factor decreasing from 0.90 ± 0.02 at the first attended location to 0.55 ± 0.05 at the 8th attended location: While the model was attending to locations with high SFC, these were not necessarily the locations with highest SFC. It consequently seems that saliency is more than just a measure of local SFC. The model, which implements within-feature spatial competition captured subjective saliency better than the purely local SFC measure.

The SFC measure was used as a means to more quantitatively compare the performance of the model to humans. Three databases of images were used: the first database was composed of 90 color images acquired by a video camera mounted on a vehicle driving on German roads. The camera was oriented towards the front-right direction, in order to capture most traffic signs on the passenger side of the road. The second database consisted of 104 color natural scenes (outdoor and indoor) in which a red aluminum can was a salient target, and the third database consisted of 64 color natural scenes (outdoor and indoor) in which a vehicle's emergency triangle was a salient target. For each image and irrespectively of the nature of the attended locations (target or non-target), ratios of the SFC at a given attended location to the average or maximum of the SFC map were computed (**Fig. 3.11**). Whether or not the model was attending to targets, it was generally attending to locations that we would subjectively consider salient (e.g., a blue traffic sign on a green background, or a strong edge between a tree and the sky).

The SFC at attended locations was found to be significantly higher (as a population, for each database) than the average SFC, in agreement with human experiments (Reinagel & Zador, 1999). Although this result does not necessarily indicate similarity between human eye fixations and the model's attentional trajectories, it indicates that the model, like humans, is attracted to "informative" image locations, according to the common assumption

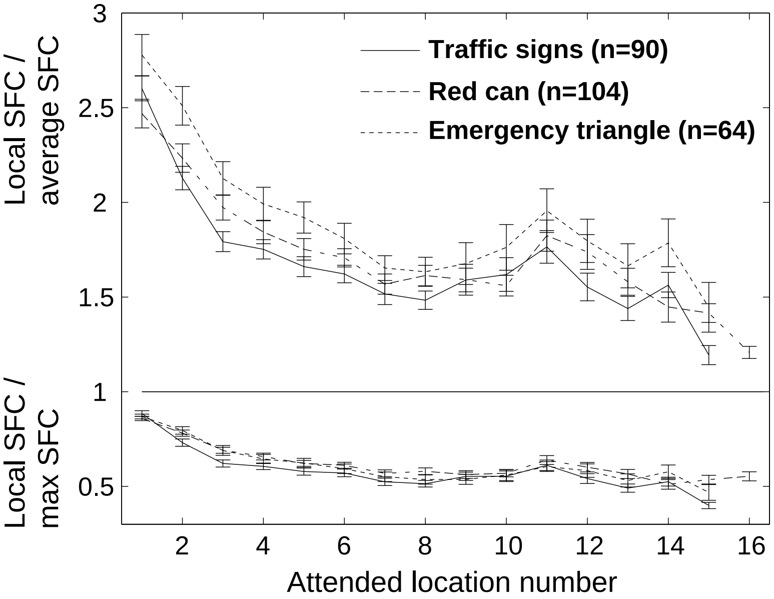


Figure 3.11: Ratios (mean ± standard error) of the spatial frequency content (SFC) locally evaluated at attended locations to the average SFC (top curves) and maximum SFC (bottom curves). In order to account for possible small positional differences between peaks in SFC and in saliency (see, e.g., **Fig. 3.10.1**), local SFC was computed as the maximum SFC over a 3×3 neighborhood, in the SFC map, of the attended location. While the local SFC was significantly higher at attended locations than on average, it was significantly lower than the maximum SFC.

that regions with richer spectral content are more informative. This result is coherent with the fact that the preattentive feature extraction mechanisms, based on center-surround operations at several spatial scales, may be viewed as a bank of spatially bandpass filters. However, although there was good agreement between the SFC map and the saliency map in many images (see for example the second row of **Fig. 3.10**), and although our SFC measure was richer than a simple analysis of the luminance channel, we found substantial disagreement between the SFC map and the saliency map for certain classes of images. These include in particular images with strong variations of illumination or color over large spatial extents: while such areas exhibit uniformly high SFC, they have low salience because their large extend does not make them particularly unique or conspicuous (**Fig. 3.10**, first, third and fourth rows). In such images, the saliency map was usually in better agreement with our subjective impression of which locations were the most conspicuous (see for example **Fig. 3.10**, first row). This result was quantitatively confirmed by the fact that, for the three image databases, local SFC at attended locations was significantly lower than the maximum SFC (**Fig. 3.11**). While the model was attending to locations with high SFC, these were not necessarily the locations with highest SFC. It consequently seems that salience is more than just a measure of local SFC. In particular, the model implements, through $\mathcal{N}(.)$, within-feature, spatial competition of candidate conspicuous locations for salience; this represents a major advantage compared to the SFC measure, which is purely local.

3.6.3 Comparison to Human Search Time with Complex Natural Images

We propose a further test in which the model's target detection performance is evaluated using a database of complex natural images, each containing a military vehicle (the "target"). Contrary to the previous studies presented above, which used low-resolution image databases with relatively large targets (typically about 1/10th the width of the visual scene), this study uses very-high resolution images (6144×4096 pixels), in which targets appear very small (typically 1/100th the width of the image). In addition, in the present study, search time is compared between the model's predictions and the average measured search times from 62 normal human observers (Toet *et al.*, 1998).

The 44 original photographs were taken during a DISSTAF (Distributed Interactive Simulation, Search and Target Acquisition Fidelity) field test in Fort Hunter Liggett, California, and were provided to us, along with all human data, by the TNO Human Factors Research Institute in the Netherlands (Toet *et al.*, 1998). The field of view for each image is $6.9 \times 4.6°$. Each scene contained one of nine possible military vehicles, at a distance ranging from 860 to 5822 meters from the observer. Each slide was digitized at 6144×4096 pixels resolution. Sixty-two human observers aged between 18 and 45 years and with visual acuity better than 1.25 arcmin^{-1} participated to the experiment (about half were women and half men). Subjects were first presented with 3 close-up views of each of the 9 possible target vehicles, followed by a test run of 10 trials. A Latin square design (Wagenaar, 1969) was then used for the randomized presentation of the images. The slides were projected such that they subtended $65 \times 46°$ visual angle to the observers (corresponding to a linear magnification by about a factor 10 compared to the original scenery). During each trial, observers pressed a button as soon as they had detected the target, and subsequently indicated at which location on a 10×10 projected grid they had found the target. Further details on these experiments can be found in (Toet *et al.*, 1998; Bijl *et al.*, 1997).

The model was presented with each image at full resolution. Contrary to the human experiment, no close-ups or test trials were presented to the model. The most generic form of the model described above was used, without any specific parameter adjustment for this experiment. Simulations for up to 10,000 ms of simulated time (about 200-400 attentional shifts) were done on a Digital Equipment Alpha 500 workstation. With these high-resolution images, the model comprised about 300 million simulated neurons. Each image was processed in about 15 minutes with a peak memory usage of 484 megabytes (for comparison, a 640×480 scene was typically processed in 10 seconds, and processing time approximately scaled linearly with the number of pixels). The focus of attention (FOA) was represented by a disk of radius 340 pixels **(Figs. 3.12, 3.13** and **3.14)**. Full coverage of the image by the FOA would hence require 123 shifts (with overlap); a random search would thus be expected to find the target after 61.5 shifts on average. The target was considered detected when the focus of attention intersected a binary mask representing the outline of the target, which was provided with the images. Three examples of scenes and model trajectories are presented in **Figs. 3.12, 3.13** and **3.14**; In the first image, the target was immediately found by the model, while, in the second, a serial search was necessary before the target could be found, and, in the third, the model failed to find it.

The model immediately found the target (first attended location) in seven of the 44 images. It quickly found the target (fewer than 20 shifts) in another 23 images. It found the target after more than 20 shifts in 11 images, and failed to find the target in three images. Overall, the model consequently performed surprisingly well, with a number of attentional shifts far below the expected 61.5 shifts of a random search in all but six images. In these six images, the target was extremely small (and hence not conspicuous at all), and the model cycled through a number of more salient locations.

The following analysis was performed to generate the plot presented in **Fig. 3.15**: First, a few outlier images were discarded, when either the model did not find the target within 2000 ms of simulated time (about 40-80 shifts; six images), or when half or more of the humans failed to find the target (three images), for a total of eight discarded images. An average of three overt shifts per second was assumed for the model, hence allowing us to scale the model's simulated time to real time. An additional 1.5 second was then added

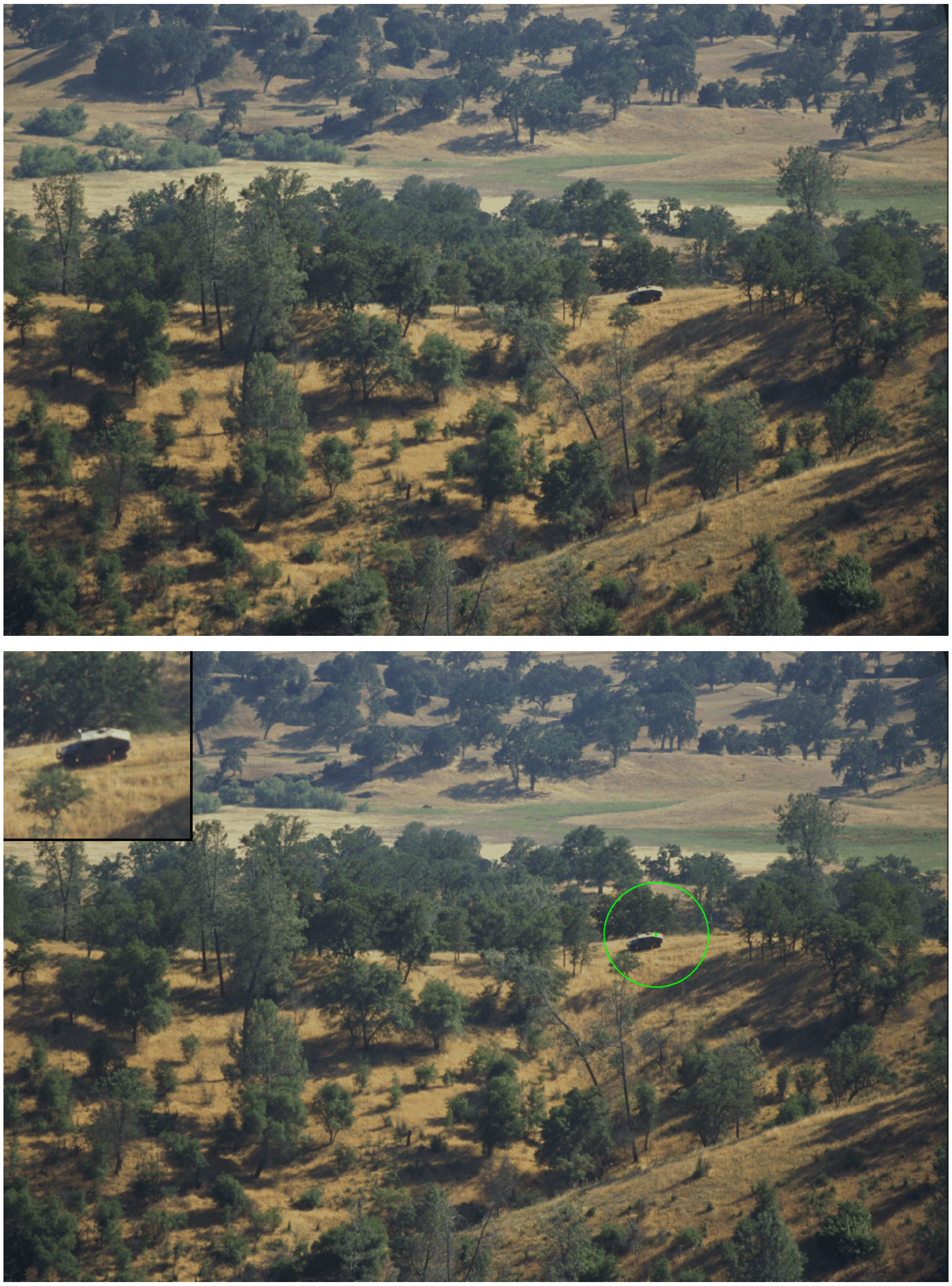


Figure 3.12: Example of image from the database of 44 scenes depicting a military vehicle in a rural background. The algorithm operated on 24-bit color versions of these 6144×4096 pixel images and took on the order of 15 min real time on Dec Alpha workstation to carry out the saliency computation. **(Top)** Original image; humans found the location of the vehicle in 2.6 sec. on average. **(Bottom)** The vehicle was determined to be the most salient object in the image, and was attended first by the model. Such a result indicates strong performance of the algorithm in terms of artificial vision using complex natural color scenes. After scaling of the model's simulated time such that it scans two to four locations per second on average, and adding a 1.5 sec. period to account for the human's latency in motor response, the model found the target in 2.2 sec.

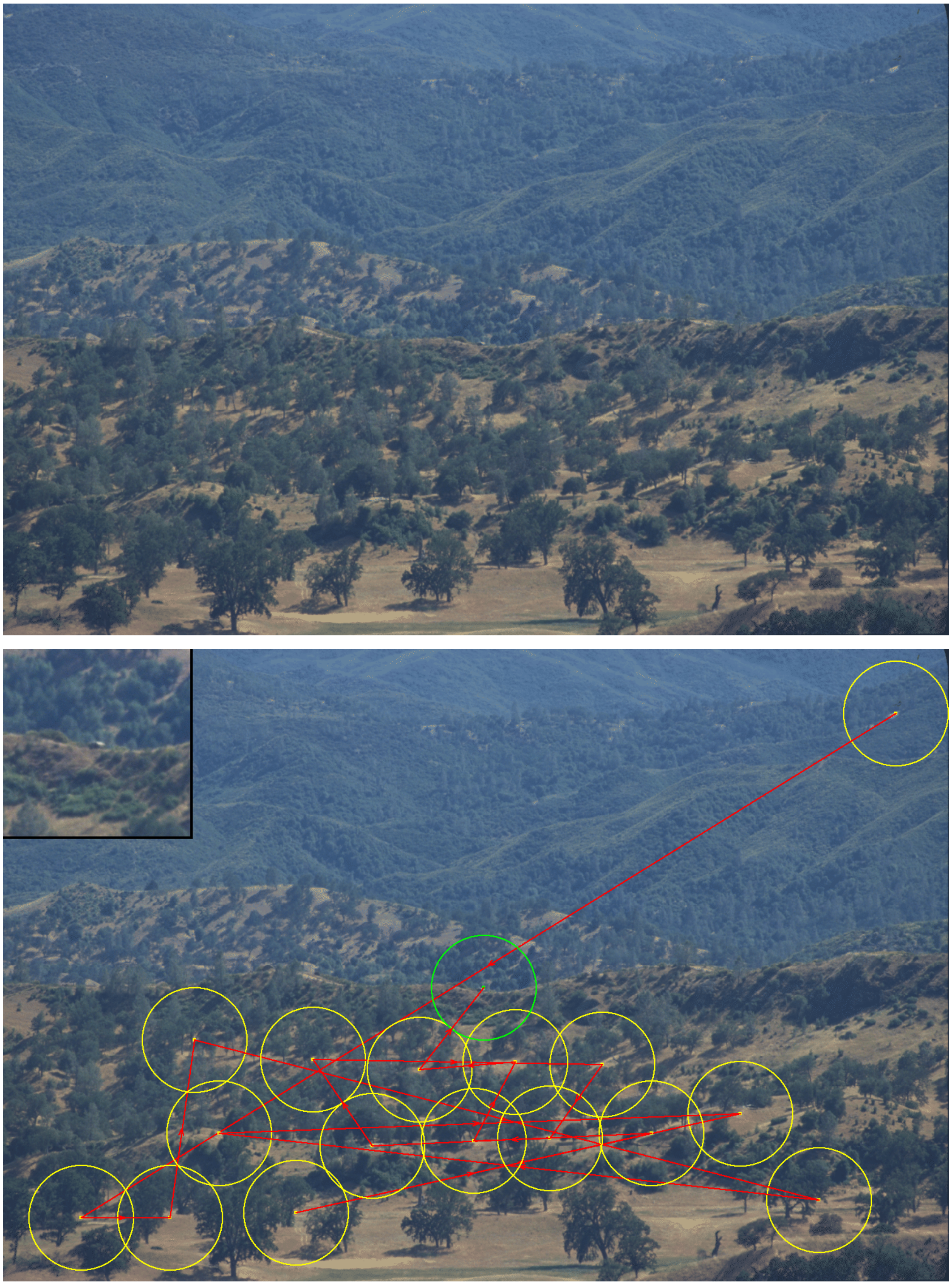


Figure 3.13: A more difficult example from the image database of military vehicles. **(Top)** The input color image. Humans found the location of the vehicle in 7.5 sec. on average. **(Bottom)** The target is not the most salient object, and the model searches the scene in order of decreasing saliency. The algorithm came to rest on the location of the target on the 17th shift, after 6.1 sec (using same time scaling as in the previous figure).

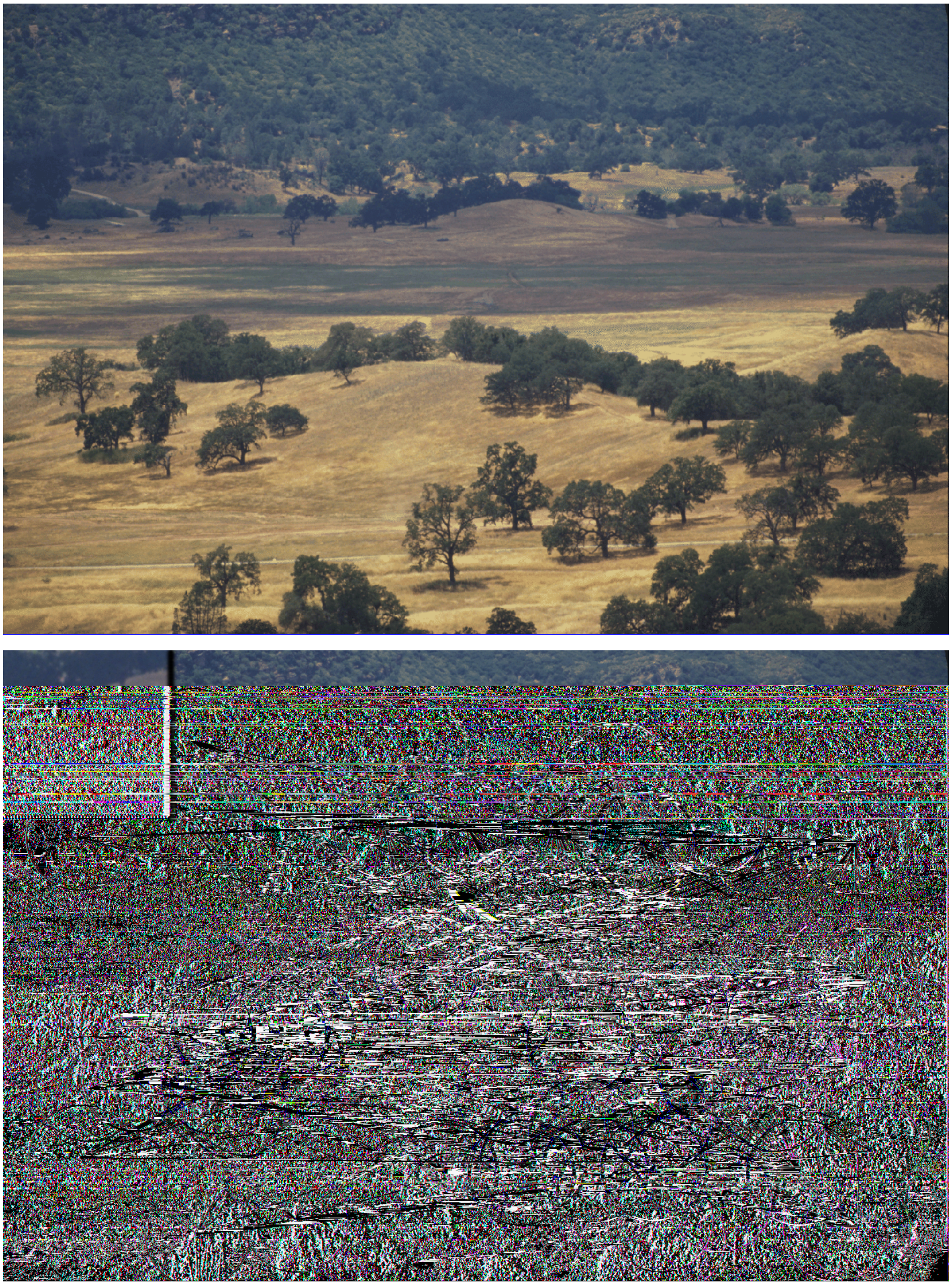


Figure 3.14: An example of model failure in the image database of military vehicles. **(Top)** The input color image. Humans found the location of the vehicle in 8.0 sec. on average. **(b)** The target is extremely small, and not salient at all. The model hence failed to find it. Inspection of the feature maps indicated that the target yielded responses in the different feature dimensions which are very similar to other parts of the image (foliage and trees). This image is one of the eight which was excluded from our analysis in **Fig. 3.15** because either humans or the model failed to reliably find the target.

to the model time to account for human motor response time. With such calibration, the fastest reaction times for both model and humans were approximately 2 seconds, and the slowest approximately 15 seconds, for the 36 images analyzed.

The results plotted in **Fig. 3.15** overall show a poor correlation between human and model search times. Surprisingly, however, the model appeared to find the target faster than humans in 3/4 of the images (points below the diagonal), despite the rather conservative scaling factors used to compare model to human time. In order to make the model's performance equal (on average) to that of humans, one would have to assume that humans shifted their gaze not faster than twice per second, which seems unrealistically slow under the circumstances of a speeded search task on a stationary, non-masked scene. Even if eye movements were that slow, most probably would humans still shift covert attention at a much faster rate between two overt fixations.

3.6.4 Comparison to Human Attention

Our model is mostly concerned with shifts of covert attention, which, as mentioned above, may substantially differ from eye movements. A study has however recently started in our laboratory, led by Drs. Steffen Egner and Christian Scheier, which proposes to attempt to build a saliency map from human observer responses measured using either eye movements, finger pointing or mouse pointing. The motivation for measuring human responses in three different manners is to try and decouple image-based saliency, which should be independent of the response mechanism, from the particular mechanistic contingencies of each of these three response modalities.

Three classes of images are being studied: Search arrays, web pages, and natural scenes **(Fig. 3.16)**. Each image is presented for four seconds, and the observers are instructed to either look or point towards the image locations which they find attract their eyes. Data is collected using a custom system, MediaAnalyzer, developed by Drs. Egner and Scheier. In the particular case of eye movements, the system uses a consumer-electronics video camera to image one of the observer's eyes during the experiment.

While only preliminary results are available at the time of this printing, several general remarks can already be made. First, at least for the first few fixations, the responses obtained *via* the three possible response modalities are highly correlated. This directly supports the idea that, at least for the first few fixations, bottom-up processes may play a significant role at directing attention and subsequent eye or hand movements. Second, our model so far correlates reasonably well with the observers' responses, again, at least for the first few fixations.

Several shortcomings of our model have however already been pointed out by this experiment, which relate to the fact that it does not incorporate any top-down processing: For example, observers have a tendency to explore first locations which are closer to the

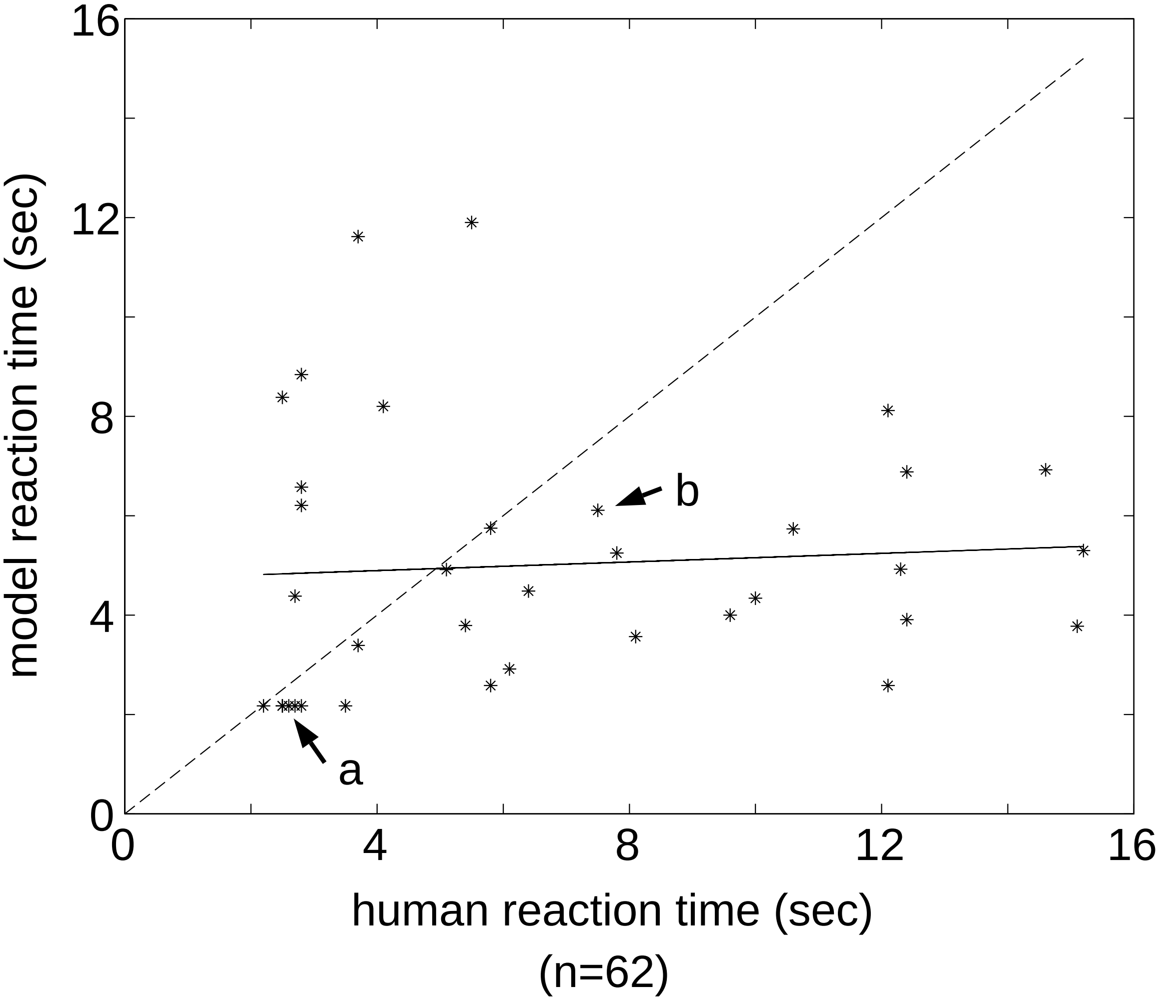


Figure 3.15: Mean reaction time to detect the military target for 62 human observers and for our deterministic algorithm. Eight of the 44 original images are not included, in which either the model or the humans failed to reliably find the target. For the 36 images studied, and using the same scaling of model time as in the previous two figures, the model was faster than humans in 75% of the images. In order to bring this performance down to 50% (equal performance for humans and model), one would have to assume that no more than two locations can be visited by the algorithm each second. Arrow **(a)** indicates the "pop-out" example of **Fig. 3.12**, and arrow **(b)** the more difficult example presented in **Fig. 3.13**.

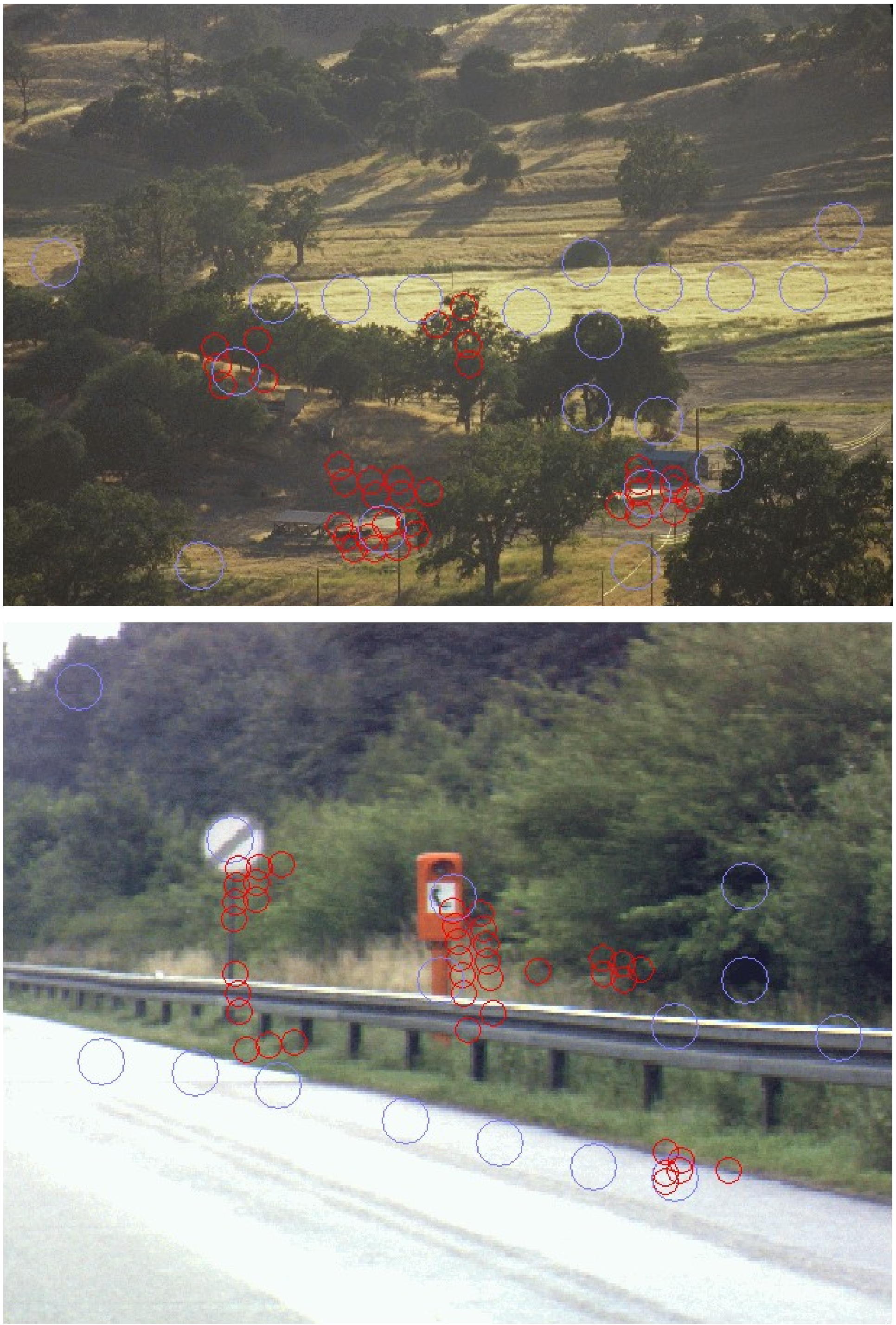


Figure 3.16: Examples of comparison to human eye movements. Eye movements were recorded using a video-based pupil tracking system. Locations of the image visited by the observer are indicated by the small red circles. The larger blue circles indicate the locations attended to by our model. The model was run for a longer time than the human recording, which explains why it visited more locations than the human observer. For the first few visited locations, however, we find an encouraging correlation between the model and human data.

fixation cross presented between each image to be analyzed. Our model at present does not incorporate any such prior bias, and does not favor any particular spatial location (beyond the selection which is made on local bottom-up saliency) when inspecting a new image. Our model also has a tendency to generate fewer fixations than humans do; this aspect of our model could easily be adjusted by playing with the inhibition-of-return and proximity preference rules already implemented. Because this Chapter, however, is mostly concerned with purely bottom-up attentional control in the absence of any top-down mechanism, we will leave such further enhancements of our model, with the goal of better matching real human attentional scanpaths by incorporating a number of top-down biases, to future studies.

3.7 Application to Image Compression

By construction of the model and its biological inspiration, we may consider that the saliency map topographically encodes for the probability for a generic observer to attend to given locations in the visual scene. While this only is a gross approximation of the behavior of such generic observer (see next Section), it however can be argued that at least the first few image locations visited by an observer are selected based on their bottom-up saliency. Using our saliency map as such probability measure provides a powerful means of spatially modulating the signal-to-noise ratio (SNR) of image compression algorithms, such that salient objects receive a high SNR while non-salient objects are compressed with high loss and low SNR.

We built a non-standard compression program exploiting this principle **(Fig. 3.17)**. The algorithm developed is based on the JPEG algorithm. In JPEG, image patches are transformed using a discrete cosine transform (DCT; a lossless operation similar to a Fourier transform), after which a variable amount of loss is introduced through quantization of the DCT's results by using a small number of possible values for each DCT coefficient. The quantized results are then further compressed using a lossless Huffman encoding. The steps in the quantizer are fixed in the JPEG standard and have been determined through psychophysical experiments. In our implementation of saliency-based compression, we spatially varied the quantizer according to the local measure of salience: At locations with high salience, a finer quantizer was used, which approximated the actual DCT coefficients in a more faithful manner. This faithful approximation also resulted in a high number of bits being necessary to encode the quantized results. Progressively coarser quantizers were used at locations with progressively lower saliency, hence resulting in progressively higher compression ratios and higher losses. Because we vary the expression of the quantizer in a spatially dependent manner, a feature which is not supported by the JPEG standard, images compressed with our algorithm cannot be viewed using standard JPEG decoders. Note finally that our algorithm requires the saliency map be included (after lossless compression

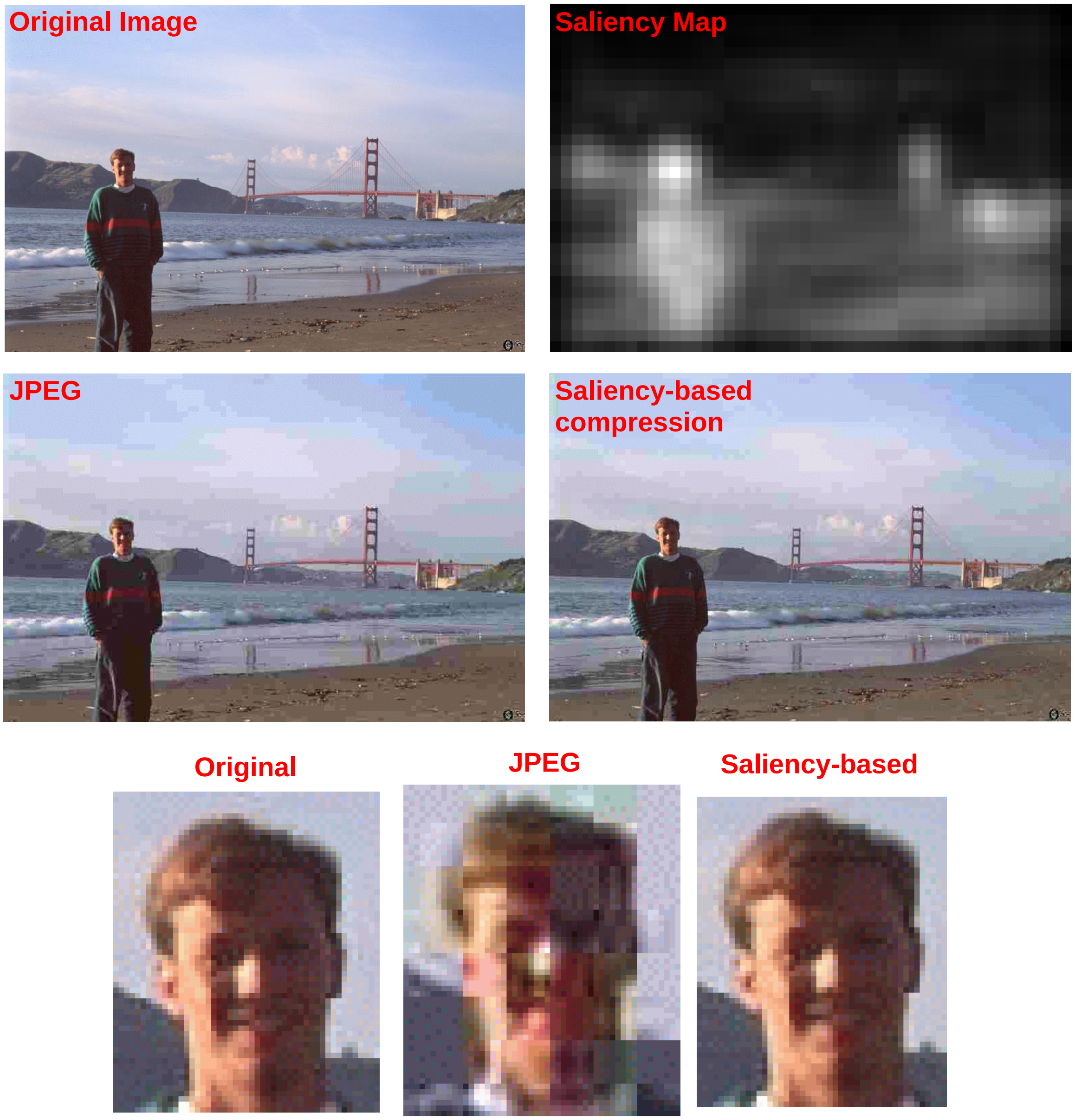


Figure 3.17: Saliency-based image compression. By spatially modulating the coarseness of the JPEG quantizer based on the local measure of saliency, our algorithm yields more faithful reproduction of salient image locations at the detriment of less salient image locations, for a given total compressed image size. In this example, the same compression ratio was achieved both by JPEG and by our algorithm (in which the saliency map is stored with the compressed image, to allow decompression). Significantly lower loss was, however, introduced by our algorithm in salient locations such as the person's face (see magnified excerpts in bottom row), at the expense of introducing higher losses in other image regions.

using a Huffman code) with the compressed image in order to allow decompression; this, however, has only a minimal impact onto the final compressed image size, because of the low spatial scale of the saliency map (which is reduced by a factor 16 in x and y compared to the original image).

With this algorithm, it is possible, for a given desired compressed image size, to devote a larger fraction of that size to salient objects than is done using JPEG. Visually, images compressed with our algorithm are more pleasant to look at, because the locations which we are likely to visit are compressed with lower loss. Note, however, that often our algorithm appears to perform worse than JPEG for the same compression ratio when using standard image quality metrics such as average SNR, simply because our method introduces higher losses in often extended non-salient image regions in order to be able to preserve the often smaller salient locations with higher fidelity.

Saliency-based image compression would be most successful with animated sequences, in which the observer would mainly attend to salient locations and not suffer from the high losses introduced at non-salient locations, as the latter would only be perceived by coarse, extrafoveal visual mechanisms. It would finally be important when developing a commercial saliency-based image compression algorithm to ensure that no new salient location is created by compression artifacts at originally non-salient locations. This aspect currently is not accounted for by our compression algorithm. Ideally, it would require the saliency computation algorithm and the compression algorithm to work in tandem.

3.8 Application to Artistic and Advertising Design

Our model is a purely bottom-up model; it will consequently not tell exactly where a typical observer would look, but rather what are the conspicuous regions in the image. While this may seem to be a subtle difference, it is of importance, since the model may miss some features which we consider very conspicuous by training (e.g., eyes and facial features, for which humans develop special pattern-specific neurons, not implemented in the model). Nevertheless, we evaluated the appropriateness of this model to the evaluation of artistic designs, with particular emphasis onto the evaluation of our own scientific posters **(Figs. 3.18, 3.19, 3.20)** and of some magazine covers **(Fig. 3.21)**.

Potentially, the fact that the model only emulates the bottom-up part of our visual system may turn out to be even more interesting than if it were trying to fully emulate an average human observer: Indeed, we all have a fairly good idea of what an average observer would see in a given scene. The model, however, can point us to the fact that some aspects of the design might be suboptimal at grabbing an observer's attention **(Fig. 3.21)**.

For example, much of the text in **Fig. 3.21** has very poor contrast with the underlying textured dress in the second image, and was fully ignored by our model. In some images of

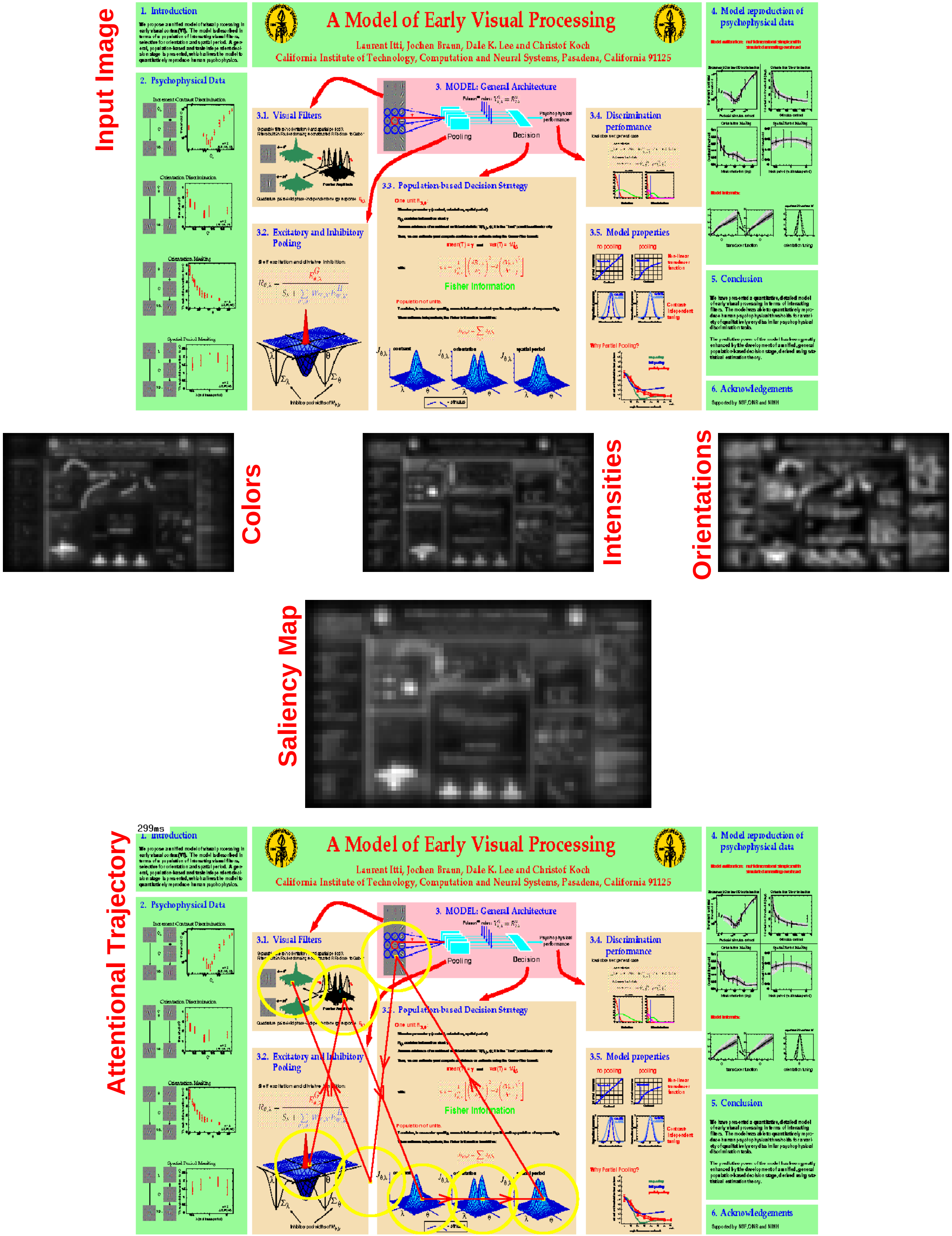

Input Image
1. Introduction
2. Psychophysical Data
A Model of Early Visual Processing
Laurent Itti, Jochen Braun, Dale K. Lee and Christof Koch
California Institute of Technology, Computation and Neural Systems, Pasadena, California 91125
3. MODEL: General Architecture
Pooling
Decision
3.1. Visual Filters
3.2. Excitatory and Inhibitory Pooling
3.3. Population-based Decision Strategy
Fisher Information
3.4. Discrimination performance
3.5. Model properties
4. Model reproduction of psychophysical data
5. Conclusion
6. Acknowledgements
Colors
Intensities
Orientations
Saliency Map
Attentional Trajectory
299ms

Figure 3.18: Initial model predictions on one of our scientific conference posters. The model is immediately attracted to several of our surface renderings because of their large size, high color contrast and high luminance contrast. Only later does the model visit the poster's overview box (towards top-center), which we wanted to be the first attended location.

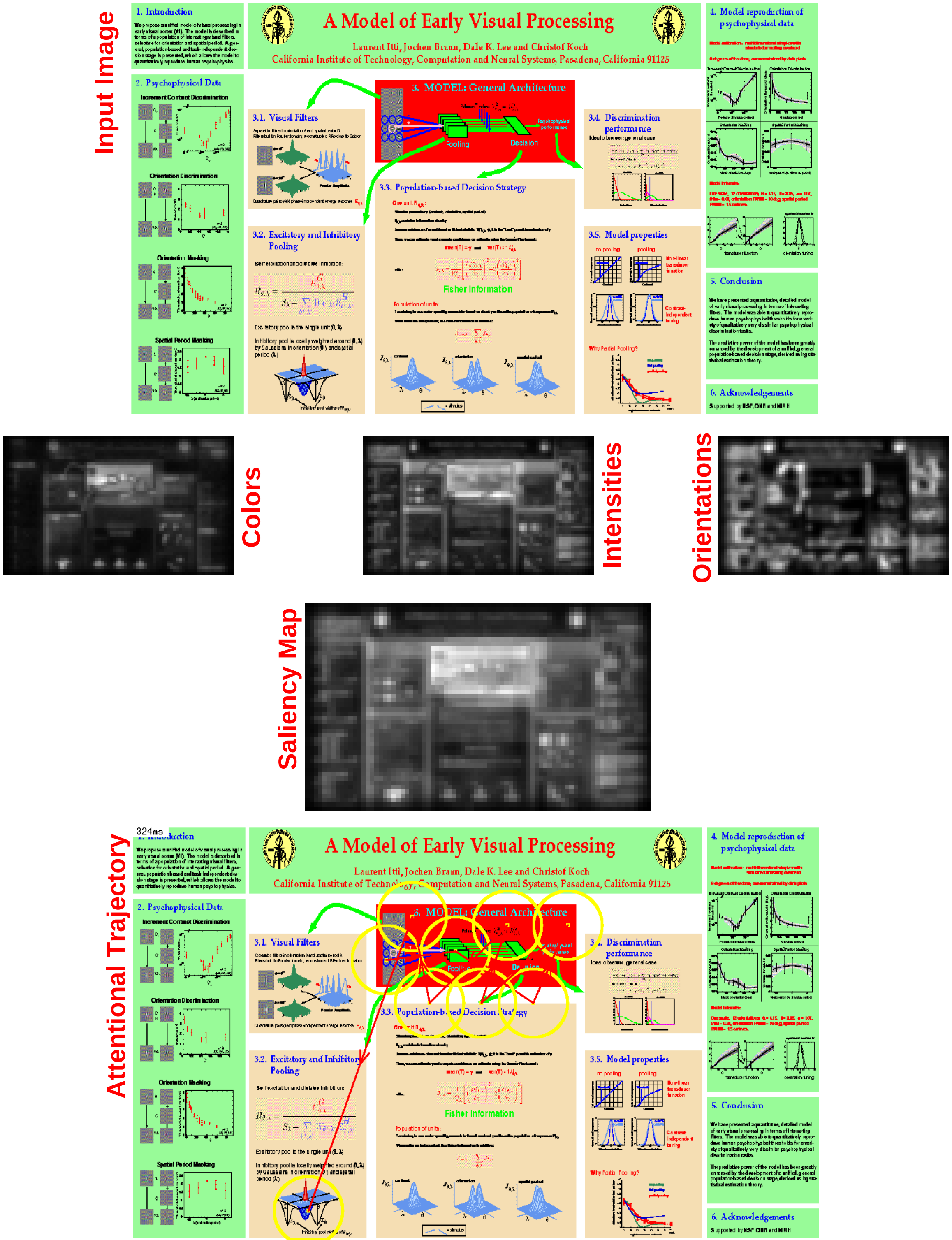

Input Image
1. Introduction
2. Psychophysical Data
Increment Contrast Discrimination
Orientation Discrimination
Orientation Masking
Spatial Period Masking
A Model of Early Visual Processing
Laurent Itti, Jochen Braun, Dale K. Lee and Christof Koch
California Institute of Technology, Computation and Neural Systems, Pasadena, California 91125
3. MODEL: General Architecture
Pooling
Decision
3.1. Visual Filters
3.2. Excitatory and Inhibitory Pooling
3.3. Population-based Decision Strategy
Fisher Information
3.4. Discrimination performance
3.5. Model properties
4. Model reproduction of psychophysical data
5. Conclusion
6. Acknowledgements
Colors
Intensities
Orientations
Saliency Map
Attentional Trajectory
324ms

Figure 3.19: Model predictions on a revised design of our poster. The initially salient surface plots have been shrunk and colored with less-contrasted hues. As a direct consequence, we can see how their response in all three conspicuity maps have been reduced compared to the previous figure. In addition, more saturated colors have been used for the overview box, which has clearly become the most salient box in the poster.

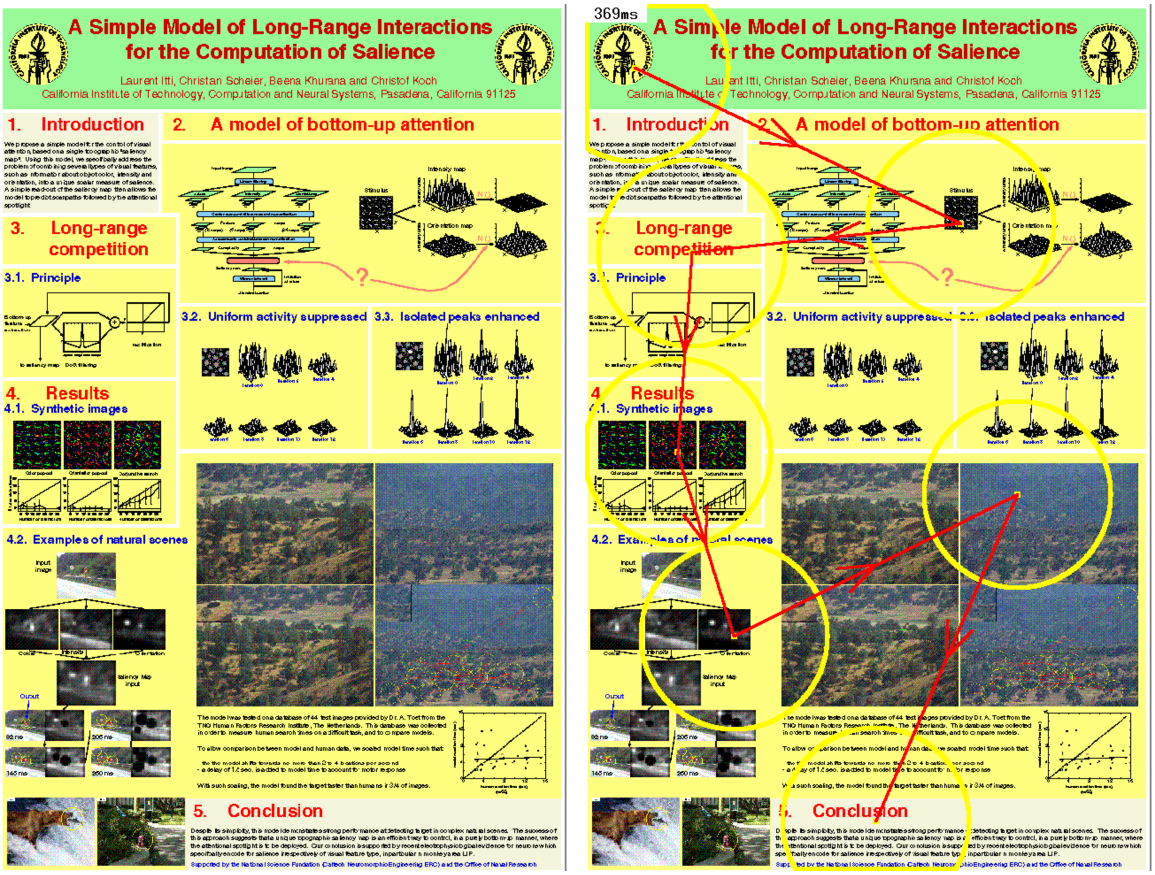

369ms
A Simple Model of Long-Range Interactions for the Computation of Salience
Laurent Itti, Christian Scheier, Beena Khurana and Christof Koch
California Institute of Technology, Computation and Neural Systems, Pasadena, California 91125
1. Introduction
2. A model of bottom-up attention
3. Long-range competition
3.1. Principle
3.2. Uniform activity suppressed
3.3. Isolated peaks enhanced
4. Results
4.1. Synthetic images
4.2. Examples of natural scenes
5. Conclusion

Figure 3.20: Model predictions on another of our posters. This time, our poster has been carefully designed with special care given to the saliency of its individual elements. Our model almost exactly follows the order in which the different sections of the poster are numbered, hence suggesting that no further design improvement is necessary.

magazine covers similar to those shown, the magazine's bar-code was selected by the model because it was determined to be similarly or more salient than other locations in the image, such as the magazine's title. In some very crowded images, the model sometimes does not find anything more interesting than some large, coarse-scale edges (e.g., of a body against a background), while such edges may not have been intended by the designer as becoming some of the most conspicuous image locations.

These findings I think are non-trivial for human observers to realize, because top-down influences naturally make us concentrate on other parts of the picture, such as a pretty face, while reviewing a given design with no time constraint. However, our model's predictions certainly indicate that a non-negligible probability exists that the first glance of some observers will go to these uninteresting (in terms of selling the magazine) locations.

3.9 Outlook

In this Chapter, we have shown that our model is capable of closely replicating human performance on a number of visual search tasks. In addition, we have explored a number of applications of the model, for example to target detection, image compression or advertising design evaluation.

We have shown that one of the key features of our model was its ability to represent simple visual features *in context*, in particular by evaluating in detail our iterative within-feature competition for salience scheme, and by comparing the model to a simple measure of spatial frequency content. In the next Chapter, we discuss these results in terms of both biological and computational vision systems.

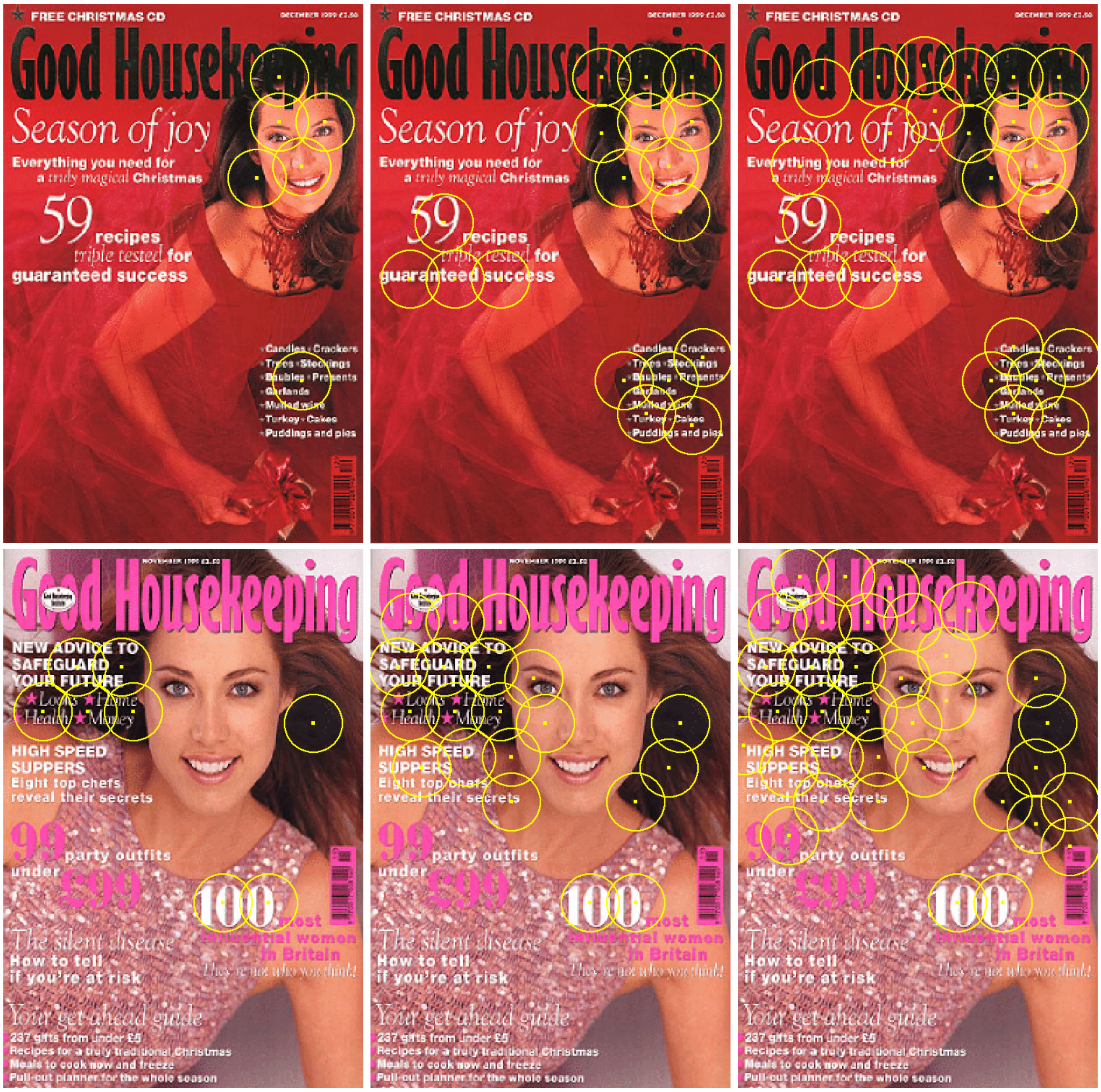

FREE CHRISTMAS CD
Good Housekeeping
Season of joy
Everything you need for a truly magical Christmas
59 recipes triple tested for guaranteed success
Candles · Crackers
Trees · Stockings
Baubles · Presents
Garlands
Mulled wine
Turkey · Cakes
Puddings and pies
Good Housekeeping
NEW ADVICE TO SAFEGUARD YOUR FUTURE
Looks ★ Home
Health ★ Money
HIGH SPEED SUPPERS
Eight top chefs reveal their secrets
99 party outfits under £99
100 most influential women in Britain
They're not who you think!
The silent disease
How to tell if you're at risk
Your get-ahead guide
237 gifts from under £5
Recipes for a truly traditional Christmas
Meals to cook now and freeze
Pull-out planner for the whole season

Figure 3.21: Model predictions on two magazine covers. Three snapshots have been taken during the model's attentional trajectory simulation for each cover. For clarity, only the focus of attention is displayed. Although we do not claim that the trajectories generated by the model would necessarily match those generated by human observers, we believe that the model can prove useful in the evaluation of these designs. For example, based purely on bottom-up cues and without any knowledge of the semantic content of these images, the model ignores large portions of the text in the second image; instead, it finds that the publication's bar-code is more salient, and visits it.

Chapter 4 Neurobiological and Computational Implications

We have demonstrated that a relatively simple processing scheme, based on some of the key organizational principles of pre-attentive early visual cortical architectures (center-surround receptive fields, non-classical within-feature inhibition, multiple maps) in conjunction with a single saliency map performs remarkably well at detecting salient targets in cluttered natural and artificial scenes.

Key properties of our model, in particular its usage of inhibition-of-return and the explicit coding of saliency independent of feature dimensions, as well as its behavior on some classical search tasks, are in good qualitative agreement with the human psychophysical literature.

It can be argued, based on the tentative scaling between simulated model time and human time described in the experiment using military images (disregarding the fact that our computer implementation required on the order of 15 minutes to converge for the 6144×4096 pixel images versus search times on the order of a 2-20 seconds for human observers, and disregarding the fact that our algorithm did not deal with the problem of detecting the target in the focus of attention), that the bottom-up saliency-based algorithm outperforms humans in a demanding but realistic target detection task involving camouflaged military vehicles.

One paradoxical explanation for this superior performance might be that top-down influences play a significant role in the deployment of attention in natural scenes. Top-down cues in humans might indeed bias the attentional shifts, according to the progressively constructed mental representation of the entire scene, in inappropriate ways. Our model lacks any high-level knowledge of the world and operates in a purely bottom-up manner.

This does suggest that for certain (possibly limited) scenarios, such high-level knowledge might interfere with optimal performance. For instance, human observers are frequently tempted to follow roads or other structures, or may "consciously" decide to thoroughly examine the surroundings of salient buildings that have popped-out, while the vehicle might be in the middle of a field or in a forest.

4.1 Computational Implications

The main difficulty we encountered was that of combining information from numerous feature maps into a unique scalar saliency map. Most of the results described above do not hold for intuitively simple feature combination schemes, such as straight summation. In particular, straight summation fails to reliably detect pop-outs in search arrays such as those shown in **Fig. 3.4**. The reason for this failure is that almost all feature maps contain numerous strong responses (e.g., the intensity maps show strong activity at all target and distractor elements, because of their high contrast with the black background); the target consequently has a very low signal-to-noise ratio when all maps are simply summed. Here, we proposed a novel solution, which finds direct support in the human and animal studies of non-classical receptive-field interactions.

The first computational implication of our model is that a simple, purely bottom-up mechanism performs surprisingly well on real data in the absence of task-dependent feedback. This is in direct contrast to some of the previous models of visual search, in which top-down bias was almost entirely responsible for the relative weighting between the feature types used (Wolfe, 1994).

Further, although we have implemented the early feature extraction mechanisms in a comparatively crude manner (e.g., by approximating center-surround receptive fields by simple pixel differences between a coarse and a fine scale versions of the image), the model demonstrates a surprising level of robustness, which allows it to perform in a realistic manner on many complex natural images. We have studied the robustness of a pop-out signal in the presence of various amounts of added speckle noise, and have found that the model is almost entirely insensitive to noise as long as such noise is not directly masking the main feature of the target in spatial frequency or chromatic frequency space. We believe that such robustness is another consequence of the within-feature iterative scheme which we use to allow for the fusion of information from several dissimilar sources. This result also represents a simple prediction of the model which could be tested in human observers.

That our model yields robust performance on natural scenes is not too surprising when considering the evidence from a number of state-of-the-art object recognition algorithms (Poggio, 1997; Niyogi *et al.*, 1998; Malik & Perona, 1990; Simoncelli *et al.*, 1992). Many of these demonstrate superior performance when compared to classical image processing schemes, although these new algorithms are based on very simple feature detection filters, similar to the ones found in biological systems.

4.2 Neurobiological Implications

While our model reproduces certain aspects of human search performance in a qualitative fashion, a more quantitative comparison is premature for several reasons.

Firstly, we have yet to incorporate a number of known features. For instance, we did not include any measure of saliency based on temporal stimulus onset or disappearance, or on motion (Hillstrom & Yantis, 1994). We also have not yet integrated any retinal non-uniform sampling of the input images, although this is likely to strongly alter the saliency of peripherally-viewed targets (but see **Fig. 4.1** for a prototype). The main reason for not using such pre-processing in our simulations is that it renders the model less invariant with respect to spatial scale, as the non-uniform filtering process implies a given viewing distance. In a hardware implementation where a fixed set of cameras are used, it has however been shown by Toepfer *et al.* (1998) that using such representation can significantly accelerate processing when all of the linear filtering operations are properly adapted to the new complex logarithmic input space. In our model, we expect that using such biologically-plausible retinal processing stage would yield more realistic attentional scanpaths, particularly for shifts of overt attention.

We also have not yet addressed the detailed timing differences found in the well-known asymmetries in search tasks (Treisman & Gormican, 1988), although our model appeared to reproduce the basic finding of an asymmetry. Spatial "grouping" acting among stimuli is also known to dramatically affect search time performance (Driver *et al.*, 1992) and has not been dealt with here. In principle, this can be addressed by incorporating excitatory, cooperative center-surround interactions among neurons both within and across feature maps. And, as discussed above, our model is completely oblivious to any high-level features in natural scenes, including social cues.

More importantly, a number of electrophysiological findings muddy the simple architecture our model operates under **(Fig. 1b)**. Single-unit recordings in the visual system of the macaque indicate the existence of a number of distinct maps of the visual environment that appear to encode the saliency and/or the behavioral significance of targets. These include neurons in the superior colliculus, the inferior and lateral subdivisions of the pulvinar, the frontal-eye fields and areas within the intraparietal sulcus (Laberge & Buchsbaum, 1990; Robinson & Petersen, 1992; Kustov & Robinson, 1996; Gottlieb *et al.*, 1998; Colby & Goldberg, 1999). What remains unclear is whether these different maps emphasize saliency for different behaviors or for different visuo-motor response patterns (for instance, for attentional shifts, eye or hand movements). If saliency is indeed encoded across multiple maps, this raises the question of how competition can act across these maps to ensure that only a single location is chosen as the next target of an attentional or eye shift.

Following Koch and Ullman's (Koch & Ullman, 1985) original proposal that visual search is guided by the output of a selection mechanism operating on a saliency map, it now seems plausible that such a process does characterize processing in the entire visual system. Inhibition-of-return (IOR) is a critical component of such search strategy, which essentially acts as memory. If its duration is reduced, the algorithm fails to find less salient objects because it endlessly cycles through the same number of more salient objects. For instance, if the time scale of IOR was reduced from 900 ms to 50 ms, the model would detect the most salient object, inhibit its location, then shift to the second most salient location, but it would subsequently come back to the most salient object, whose inhibition would have ceased during the attentional shift from first to second object. Under such conditions, the algorithm would never focus on anything else than the two most salient locations in the image. Our finding that IOR plays a critical role in purely bottom-up search may not necessarily disagree with recently suggested evidence that humans appear to use little or no memory during search (Horowitz & Wolfe, 1998); while these authors do not refute

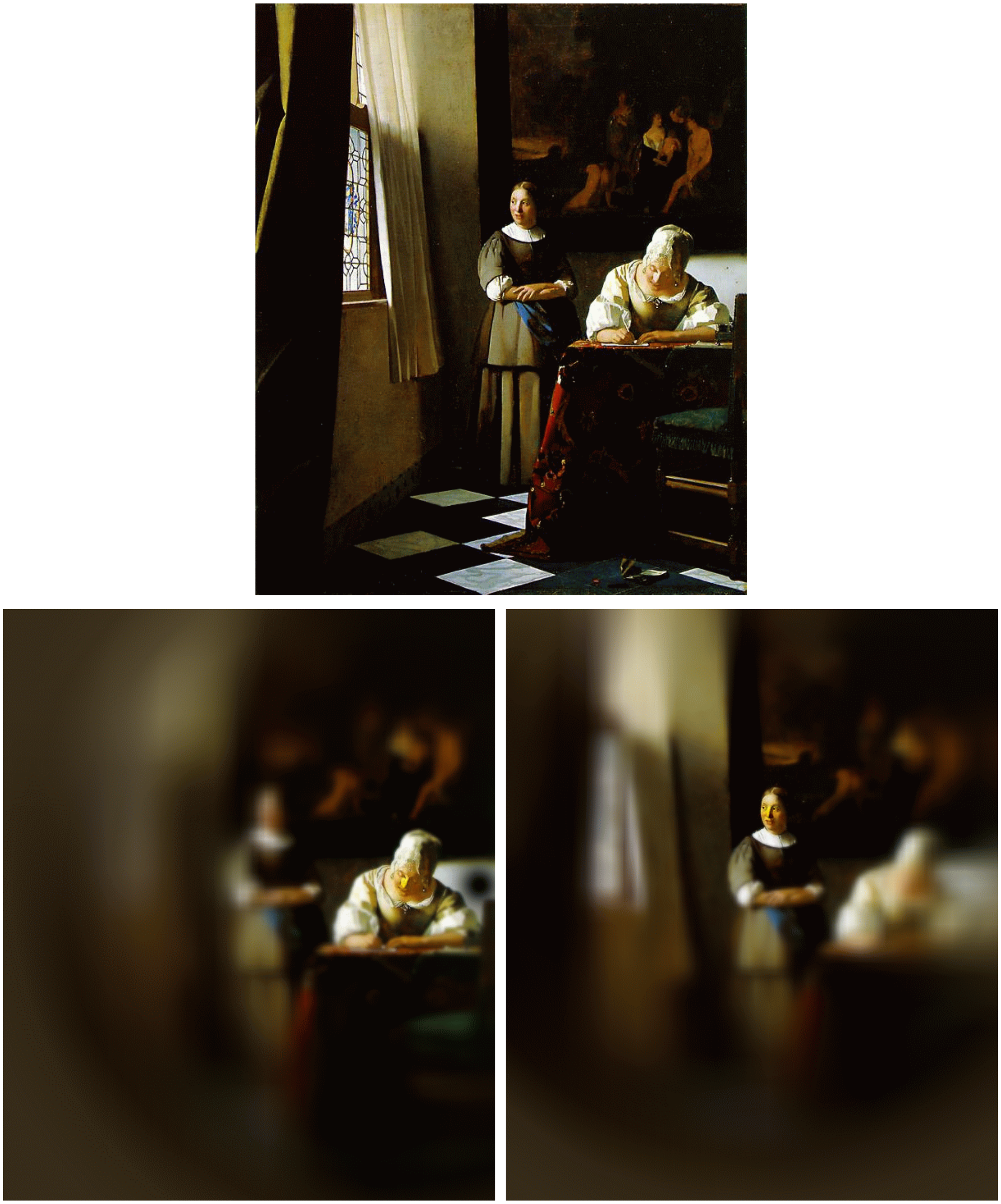


Figure 4.1: Non-uniform sampling by a retina. We implemented a simple input filter which replicates some of the features of the human retina (resolution fall-off with eccentricity, no blue cones in the fovea, blind spot at about 15° eccentricity). Two transformed versions of the original image on top are shown in the bottom row, for two different fixations.

the existence of IOR, a precise understanding of how bottom-up and top-down aspects of attention interact in human visual search remains to be elucidated.

Whether or not this implies that saliency is expressed explicitly in one or more visual field maps remains an open question. If saliency is encoded (relatively) independently of stimulus dimensions, we might be able to achieve a dissociation between stimulus attributes and stimulus saliency. For instance, appropriate visual masks might prevent the attributes of a visual stimulus to be read out without affecting its saliency. Or we might be able to directly influence such maps, for instance using reversible pharmacological techniques in animals or transcranial magnetic stimulations in human volunteers (TMS)?

Alternatively, it is possible that stimulus saliency is not expressed independently of feature dimensions but is encoded implicitly within each specific feature map as proposed by Desimone and Duncan (Desimone & Duncan, 1995). This raises the question of how interactions among all of these maps gives rise to the observed behavior of the system for natural scenes. Such an alternative has not yet been analyzed in depth by computational work (see, however, (Hamker, 1999)).

Mounting psychophysical, electrophysiological, clinical and functional imaging evidence (Shepherd *et al.*, 1986; Andersen *et al.*, 1990; Sheliga *et al.*, 1994; Kustov & Robinson, 1996; Corbetta, 1998; Colby & Goldberg, 1999) strongly implies that the neuronal structures underlying the selection and the expression of shifts in spatial attention and occulomotor processing are tightly linked. These areas include the deeper parts of the superior colliculus; parts of the pulvinar; the frontal eye fields in the macaque and its homologue in humans, the precentral gyrus; and areas in the intraparietal sulcus in the macaque and around the intraparietal and postcentral sulci and adjacent gyri in humans.

The close relationship between areas active during covert and during overt shifts of attention raises the issue of how information in these maps is integrated across saccades, in particular given the usage of both retinal and occulo-motor coordinate systems in the different neuronal maps (see, for instance, (Andersen, 1997)). This is an obvious question that will be explored by us in future computational work.

4.3 Outlook

We can now wonder about the relationship between the saliency mechanism, the top-down volitional attentional selection process, and awareness. We will describe in the next Part of this thesis a quantitative account of the action of spatial attention on various psychophysical thresholds for pattern discrimination, in terms of a strengthening of cooperative and competitive interactions among early visual filters. How can such a scheme be combined with the current selection process based on purely bottom-up sensory data? Several possibilities come to mind. First, both processes might operate independently and both mediate access to visual awareness. Computationally, this can be implemented in a straightforward manner. Second, however, top-down attention might also directly interact with the single saliency map, for instance by influencing its constitutive elements via appropriate synaptic input. If the inhibition-of-return could be selectively inactivated at locations selected under volitional control, for example by shunting (Koch, 1998), then the winner-take-all and the attentional focus would remain at that location, ignoring for a while surrounding salient objects. Although such feedback to the saliency map seems plausible and is functionally useful, it certainly does not constitute all of the top-down attentional modulation of spatial vision (Lee *et al.*, 1999a). Finally, independent saliency maps could operate for the different feature maps and both saliency and volitional forms of attention could access them independently. Current experimental evidence does not allow us to unambiguously choose among these possibilities.

Part III

Experiments and Modeling on Human Pattern Vision and its Modulation by Top-Down Attention

Chapter 5 Introduction

Perceptual thresholds for stimulus contrast, orientation, and spatial frequency have been studied for several decades (Nachmias & Sansbury, 1974; Wilson, 1980a; Legge & Foley, 1980; Watt & Morgan, 1985). Quantitative accounts of these thresholds have become increasingly refined and usually involve a population of "noisy filters" tuned to different orientations and spatial frequencies. Although earlier models postulated filters that are independent of each other, there are serious shortcomings to this approach (Bowne, 1990; Wilson & Wilkinson, 1997). More recent models postulate an interaction between filters with spatially overlapping receptive fields (Wilson & Humanski, 1993; Foley, 1994; Zenger & Sagi, 1996), specifically, the normalization of individual filter responses relative to the total response of the local filter population ("divisive inhibition" (Carandini & Heeger, 1994)). This normalization accounts naturally for several otherwise puzzling observations, among them the initial decrease and later increase of contrast discrimination thresholds with increasing stimulus contrast (Foley, 1994; Zenger & Sagi, 1996) ("dipper function" (Nachmias & Sansbury, 1974; Legge & Foley, 1980)) and the relative constancy of orientation and spatial frequency thresholds over a wide range of stimulus contrasts (Bowne, 1990; Itti *et al.*, 1998a).

An intriguing parallel to these perceptual accounts can be found in certain models of visual cortical responses to stimulus contrast and orientation (Ben-Yishai *et al.*, 1995; Somers *et al.*, 1995; Carandini *et al.*, 1997). Despite marked differences in detail, the models in question consider a population of neurons with overlapping receptive fields, broadly tuned to a range of different orientations, and normalize individual responses relative to the population response. The normalization, which in some cases is implemented as a divisive inhibition, sharpens orientation tuning (Somers *et al.*, 1995) and renders it less dependent on stimulus contrast (Carandini *et al.*, 1997). Thus, response normalization may account for both perceptual and neuronal sensitivity to contrast and orientation.

5.1 Computational Models of Spatial Vision and Orientation Tuning

Recently, strikingly similar models have been formulated for psychophysical measurements of spatial vision thresholds and for the dependence of single-unit responses in striate cortex on simple stimulus dimensions such as contrast, orientation and spatial period. A key idea that emerged in both domains is that of "divisive inhibition" between differently tuned filters (neurons) (Heeger, 1992; Heeger, 1993; Heeger *et al.*, 1996).

In the early 1980s, several computational models were proposed as a unified account for spatial vision thresholds (Legge & Foley, 1980; Phillips & Wilson, 1984). These models consisted of linear 2D filters tuned for various orientations and spatial periods, followed by a non-linear transducer function and a decision stage which compares one or more responses to a decision criterion. Recently, several models incorporating divisive inhibition between differently tuned filters have demonstrated significantly better reproduction of psychophysical data, in particular with respect to the observation that orientation tuning of the filters appears to be relatively independent of stimulus contrast (Foley, 1994; Wilson, 1993; Zenger & Sagi, 1996).

Models of orientation (or direction) tuning of neurons in striate cortex must contend with two seemingly contradictory facts: Extracellular blockage of inhibition reduces or eliminates tuning (Sillito, 1979; Nelson, 1991) while intracellular recordings do not show the expected difference in the tuning of inhibitory and excitatory inputs (Ferster, 1987) nor any effect on tuning of intracellular blockage of inhibition (Somers *et al.*, 1995). These facts are reconciled by the notion of "recurrent excitation" within a cortical column (a columnar volume of cortex containing cells of similar tuning) (Douglas *et al.*, 1991; Douglas *et al.*, 1995a; Somers *et al.*, 1995). In this view, geniculate input is responsible only for relatively broad tuning (e.g., 120° Full Width at Half Maximum (FWHM) orientation tuning) but recurrent excitation coupled with mutual inhibition between nearby columns significantly sharpens tuning (e.g., 40° FWHM orientation tuning). Although inhibition between nearby (but differently tuned) columns is critical, it is masked in intracellular recordings by massive recurrent excitation (and inhibition) from the same column, accounting for the various findings mentioned above. Interestingly, detailed computational models (Somers *et al.*, 1995; Ben-Yishai *et al.*, 1995; Suarez *et al.*, 1995) suggest that the combination of mutual inhibition and recurrent excitation is well approximated by the "divisive inhibition" idea. (The biophysical mechanism originally proposed for divisive inhibition — shunting inhibition near the soma — is not supported by the data (e.g., (Ferster & Jagadeesh, 1992)).

The composition of the "inhibitory pool," that is, of the population of filters which interact in an inhibitory manner, is somewhat controversial. Heeger (1992) and Foley (1994) equally include filters of all orientations, which implies that even orthogonally oriented stimulus components interact with each other ("cross-orientation inhibition"). However, neither contrast masking data (Phillips & Wilson, 1984; Zenger & Sagi, 1996) nor intracellular recording (Ferster, 1987; Douglas *et al.*, 1991) show evidence of cross-orientation inhibition. Zenger and Sagi (1996) assume that the inhibitory pool of a filter tuned to 0° consists of filters tuned for ±45°. Models of neurons in striate cortex postulate weighted "near-orientation inhibition," that is, a Gaussian inhibitory function centered at 0° and with a half-width of about ±60° (Somers *et al.*, 1995). Near-orientation inhibition is an attractive possibility for anatomical reasons, since it could be accomplished by indiscriminate short-range connectivity (see below).

Thus the "consensus model" emerging from psychophysics and single-cell recording comprises the following components: (i) linear filters (neuronal receptive fields) broadly tuned to orientation and spatial period; (ii) amplification of each filter (neuronal) response by self-excitation; (iii) mutual inhibition between filters (neurons) with similar tuning. The perceptual function of such a circuit would be to both sharpen filter (neuronal) tuning and to render it relatively independent of stimulus contrast.

5.2 Short-Range Connections in Striate Cortex

Striate cortex is organized in such a way that adjacent volumes of cortex process nearby regions of visual space and similar stimulus orientations (Hubel & Wiesel, 1962). Injection of tracer substances reveal a rich variety of connections (Fitzpatrick *et al.*, 1985; Blasdel *et al.*, 1985). A useful distinction can be made between short-range connections, mediating interactions between neurons with overlapping receptive fields and spanning cortical distances smaller than 1mm, and long-range connections, mediating interactions between neurons with non-overlapping receptive fields and spanning cortical distances of several mm.

Short-range connections are mediated by both horizontal and vertical arborizations of excitatory (e.g., pyramidal cells) and inhibitory cells (e.g., basket cells). Short-range interactions are indiscriminate and do not respect the boundaries of functional columns (Amir *et al.*, 1993). Connections between lower and upper layers appear limited to a horizontal range of 500 μm or less (Allison *et al.*, 1995). In the cat, the longest horizontal reach among short-range connections is provided by large basket cells in layer III, whose arborizations span 1 mm or more (Kisvarday *et al.*, 1994). Synaptic terminals of large basket cells are found in columns of all orientation preferences, including similar (43% of all cells have an orientation tuning of $\pm 30^\circ$), oblique ($35\% \pm (30 \text{ to } 60)^\circ$) and orthogonal ($22\% \pm (60 \text{ to } 90)^\circ$) orientation preferences (Kisvarday *et al.*, 1994). Large basket cells are thought to play an important role in orientation and direction tuning in cat striate cortex (Kisvarday, 1992; Eysel, 1992).

In short, the organization and connectivity of striate cortex provides ample opportunity for establishing an "inhibitory pool" of neurons tuned to different orientations, certainly for differences of 60° or less, but to some extent even for differences of up to 90°.

5.3 Long-range Connections and Nonclassical Receptive Fields

Long-range connections in striate cortex are made by axonal arbors of excitatory (pyramidal) neurons in layers III and V (Rockland & Lund, 1983; Gilbert & Wiesel, 1983). Besides range, they are distinguished by their "patchy" terminations which seem to respect the boundaries of functional columns. Indeed, long-range connections link functional columns of predominantly similar stimulus selectivity (65% of cells have an orientation tuning of $\pm 22^\circ$, and 30% a tuning of $\pm 45^\circ$) (Ts'o *et al.*, 1986; Gilbert & Wiesel, 1989; Malach *et al.*, 1993; Malach, 1994). Interestingly, the average reach of long-range connections increases markedly at increasing levels of the cortical hierarchy (from 2.1 mm in area V1 to 9 mm in area 7a) (Malach *et al.*, 1993).

Combined optical imaging and electrophysiological recording shows that long-range connections have both excitatory and inhibitory effects at their target regions, the latter mediated by inhibitory interneurons (Weliky *et al.*, 1995). Both effects are limited to regions of similar orientation preference and the balance between excitation and inhibition can be altered by changing the intensity of stimulation.

Since long-range connections link neurons with non-overlapping receptive fields, they are able to mediate effects of stimuli outside the "classical receptive field." Such effects have been known for some time, but their full complexity is only recently being appreciated. In striate cortex, responses of orientation-tuned neurons are enhanced by the presence outside the classical receptive field of figure-ground boundaries (Lamme, 1995; Zipser *et al.*, 1996) and collinear stimuli forming a contour (Kapadia *et al.*, 1994), suppressed by a dense texture of similarly-oriented stimuli (Knierim & van Essen, 1992), and either enhanced or suppressed by a sinusoidal grating, depending on its contrast and orientation (Sillito *et al.*, 1995; Levitt & Lund, 1997). In the latter case, the presence of a sinusoidal grating outside the classical receptive field can completely alter orientation tuning inside the classical receptive field.

5.4 Background on Attentional Modulation Effects

In this background Section, we briefly review relevant single-unit and psychophysical studies of attention.

5.4.1 Single-Unit Studies of Attention

In the awake macaque, neuronal responses to attended stimuli can be 20% to 100% larger than to otherwise identical unattended stimuli. This has been demonstrated in visual cortical area V1 (Press *et al.*, 1994; Motter, 1993), area V2, and area V4 (Spitzer *et al.*, 1988; Motter, 1993; Motter, 1994a; Motter, 1994b; Maunsell, 1995; McAdams & Maunsell, 1996a) when the animal discriminates stimulus orientation, and in areas MT and MST when the animal discriminates the speed of stimulus motion (Treue & Maunsell, 1996). Neuronal contrast sensitivity in area V4 is, on average, 30% larger to attended than to unattended stimuli (Reynolds *et al.*, 1994a). Even spontaneous firing rates are 40% larger when attention is directed at a neuron's receptive field (Luck *et al.*, 1997). Whether neuronal responses to attended stimuli are merely enhanced (McAdams & Maunsell, 1996a) or whether they are also more sharply tuned for certain stimulus dimensions (Spitzer *et al.*, 1988) remains controversial.

The enhancement of neuronal responses often depends on the presence of unattended stimuli nearby and for this reason is thought to reflect an attentional bias in sensory interactions (Motter, 1993; Press *et al.*, 1994; Reynolds *et al.*, 1994b; Reynolds *et al.*, 1995; Luck *et al.*, 1997). In essence, stimuli at different locations appear to suppress each other's neuronal responses and attention seems to bias this competition in favor of attended stimuli (Desimone & Duncan, 1995; Maunsell, 1995). This competitive bias is encountered both between stimuli located in the same, classical, receptive field and stimuli located in non-overlapping receptive fields (Press *et al.*, 1994).

A seemingly different kind of attentional response enhancement in area V4 occurs in parallel across the visual scene as a result of the attentional selection of a stimulus attribute such as color or luminance (Motter, 1994a; Motter, 1994b) or shape (Chelazzi & Desimone, 1994). Finally, neuronal responses in area V4 may be modulated by the relative positions of receptive field and attentional focus in the manner of a gain field (Connor *et al.*, 1996; Connor *et al.*, 1997).

5.4.2 Psychophysical Studies of Attention

We concentrate on psychophysical studies that measure discriminability d' (Green & Swets, 1966), since results expressed in terms of d' are more directly comparable to those of single-unit studies (e.g., (Salzman & Newsome, 1994)) than results expressed in other terms (e.g., reaction time). When attention is manipulated by "cueing," d' for discriminating a simple shape is significantly larger at cued (more attended) than at uncued (less attended) locations (Müller & Findlay, 1987b; Downing, 1988; Nakayama & Mackeben, 1989). However, d' for other visual tasks is less sensitive to cueing. For example, d' for detecting a sudden luminance increment is little affected by cueing (Müller & Findlay, 1987a; Müller & Humphreys, 1991), and the same is true for detecting a singular element in an array of otherwise uniform elements (a situation called "pop-out;" (Nakayama & Mackeben, 1989)).

Another way to manipulate attention is to compare the concurrent discrimination of two targets (each partially attended) with that of only one target (which is fully attended). When the two discriminations concern simple shapes (e.g., letters), concurrent performance (d') is invariably lower than separate performance (e.g., (Bonnel *et al.*, 1987; Bonnel & Miller, 1994; Duncan *et al.*, 1994)). However, when the discriminations concern luminance increments, there is little or no reduction in performance (Bonnel *et al.*, 1992).

In short, attention seems to increase the discriminability of some features (e.g., letter shape) but not others (e.g., luminance increments). However, we know of no systematic comparison of attentional effects on different discrimination thresholds and of no attempt to account computationally for such effects observed on a wide variety of tasks (but see recent modeling effort by Dosher and Lu (1997) which interpreted, using a model, attentional effects on contrast thresholds as a reduction of internal noise).

Chapter 6 Human Pattern Discrimination Thresholds in the Near Periphery

6.1 Overview

This Chapter describes the new psychophysical dataset acquired in our laboratory for the present study. We here present the two main motivations for acquiring such dataset and briefly summarize our experiments and results.

First, most previous studies of contrast masking used spatially extended patterns (at least for the mask) and thus did not distinguish interactions between spatially overlapping and non-overlapping filters. To our knowledge, only two previous studies used spatially localized target and mask patterns (Gaussian half-widths $< 0.5°$) and thus specifically addressed interactions between overlapping filters (Zenger & Sagi, 1996; Foley & Chen, 1997). In both cases, target and mask patterns were presented near the fixation point (eccentricity $< 0.8°$). We used spatially localized patterns (Gaussian half-width $< 0.7°$) presented at varying locations at 4° of eccentricity. The peripheral presentation is a potentially important difference to earlier work, as foveal and peripheral vision are known to differ (Wilson, 1991; Snowden & Hess, 1992). Both the small pattern size and the peripheral presentation were intended to emphasize interactions between spatially overlapping filters, while reducing possible non-classical surround interactions. We hoped that this limitation would render our results more tractable computationally by our model which is limited to one location in visual space (see next Chapter).

Our experiments consisted of measuring simple contrast masking thresholds for pairs of overlapping Gabor patches. The contrast, orientation and spatial frequency of one of the two patches was varied systematically, and contrast threshold for the second, superposed patch was measured. Contrast masking experiments were chosen, as mentioned previously, because presumably they would allow us to characterize the cortical interactions responsible for the observed interaction between the two overlapping Gabor patches. As contrast masking studies provide no direct evidence as to the tuning widths of individual filters, we conducted two further experiments in which observers discriminated orientation or spatial frequency of target patterns in the absence of any mask. Orientation and spatial frequency discrimination thresholds reflect the tuning width of visual filters and are expected to be relatively independent of filter interactions (Burbeck & Regan, 1983; Bradley & Skottun, 1984; Bowne, 1990). As Bowne (Bowne, 1990) and Wilson (Wilson, 1991) have pointed out, the low dependence of orientation and spatial frequency thresholds on stimulus contrast provides an important constraint for visual filter models.

The second main concern of this experimental study is that acquiring a set of five distinct experiments, using consistent stimuli across experiments, might allow for simultaneous quantitative model predictions of all the data for each of several observers. Indeed, although several groups have extensively reported detailed separate simulations of more complete versions of each of our five experiments, the only attempt we know of which tried to provide a simultaneous, unified account of a wide range of different experiments was by Bowne (Bowne, 1990). Bowne failed to simultaneously replicate discrimination thresholds for increment contrast, orientation and spatial frequency, using in particular a variant of the successful model by Wilson (Wilson, 1986). After a detailed theoretical analysis of this failure, he concluded that all models of this type are incapable of such simultaneous prediction. Bowne further suggested that assuming task-dependent noise at the level of the decision process (which relates neuronal responses to psychophysical thresholds) could allow for a simultaneous prediction. Such task-dependent central noise, however, greatly reduces the value of models, since almost anything can be predicted by appropriately adjusting the amount of central noise for each task studied, and since this central noise cannot easily be measured (because the physiological correlates of the decision process are not known).

6.2 Psychophysical Experiments

We studied the discrimination of Gabor stimuli in the near periphery with respect to contrast, orientation, and spatial frequency. Seven naive observers participated to the experiment. The observers were asked to observe simple pattern stimuli briefly flashed on a computer screen, and to report what they saw by pressing one of two possible keys. Below, we detail the technical aspects of the data collection procedure.

6.2.1 Stimulus Generation

Stimuli were generated with an SGI Indigo workstation. “Color bit stealing” (Tyler, 1997) was used to reduce the minimum luminance step of the display from 1.5% to 0.2%. Screen luminance varied from 1 to 90cd/m^2 (mean 45cd/m^2), and room illumination was 5cd/m^2. Displays subtended $16 \times 13^\circ$ of visual angle for a viewing distance of 80cm.

All tasks employed a **temporal two-alternative forced-choice (2AFC)** paradigm. In this stimulus presentation method, two alternative targets are successively presented in random order, and the observer makes a comparative judgment between the two targets **(Fig. 6.1)**. For example, the observer may report whether the target with higher contrast was presented first or second. Because this method uses comparative discrimination, it greatly reduces observer bias which may buildup, with other methods, after repeated trials. For example, in a simpler “Yes/No” paradigm, where only one target is presented and the observer reports on an attribute of the target (e.g., present or absent), drifts or even discontinuities are often observed in the decision criterion used by observers: One day, observers may be more lenient and report “Yes” even when they are uncertain, while another day, they may report “Yes” only when having high confidence about their perception. A more formal treatment of this issue is presented in the next Chapter, where we compare in computational terms the 2AFC and the “Yes/No” paradigms, within the framework of

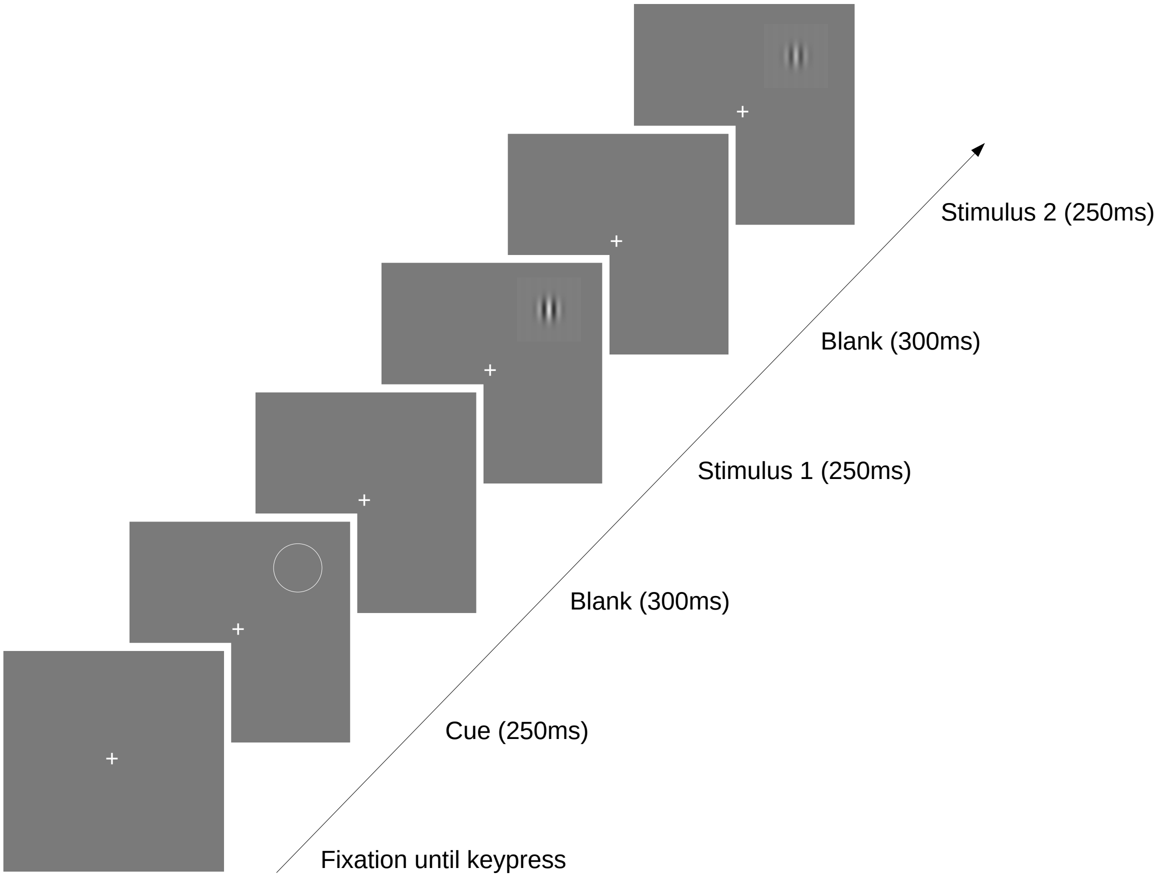


Figure 6.1: Stimulus presentation and sequencing for the experiments described. In this example, observers had to discriminate stimulus contrast, which is the only parameter distinguishing between the two stimuli. For the trial shown, the correct answer was to press the key labeled “1” (of two possible keys, labeled “1” and “2”) since the contrast of the stimulus presented first was higher than that of the stimulus presented second.

ideal-observer decision.

Each trial consisted of the following **(Fig. 6.1)**:

1. A fixation cross was presented at the center, and remained there until the observer decided to initiate the trial by pressing a key;

2. A circular cue of 1° diameter was flashed for 250 ms around the future target location. The location of this cue (and of the subsequent stimuli) was chosen randomly along a circle, such that it always was at 4° eccentricity. The role of the spatial randomization was to avoid habituation at a specific location of the visual space, and the role of the cue was to eliminate spatial uncertainty;

3. A 300 ms blank interval followed;

4. The two alternative targets were then flashed in random order, for 250 ms each, separated by a 300 ms blank interval.

5. Observers could then give their response to the question (known to them before each series of trials), under no time pressure, by pressing one of two possible keys. Observers subsequently received auditory feedback (a beeping sound in case of error, no sound in case of correct answer), and the next trial was initiated.

The total duration of each trial hence was 1350 ms, not including the response time.

6.2.2 Staircase Procedure

In each experiment, a single stimulus parameter was varied from trial to trial. Such parameter could for example be the stimulus contrast, orientation or spatial frequency, and its nature was known to the observer. Consequently, there was no task uncertainty in our experiments. For one of the two possible targets, this parameter was fixed, and for the other, it was varied from trial to trial. During one "block" of 100 trials, only this parameter was varied from trial to trial, while all other physical characteristics of the stimulus remained identical. Acquiring several blocks of trials with identical settings allowed us to collect enough data to compute the observer's "threshold" (see below) for that setting. Each threshold computation yielded one datapoint. Repeating this procedure for a variety of stimulus settings allowed us to collect 32 datapoints in this experiment, over the course of approximately six weeks at one hour per day.

During each block of 100 trials, the alternative targets were modified following a **staircase procedure**. Such procedure consists of starting a given block of trials with two easily distinguished alternate targets; for example, if contrast is the stimulus parameter distinguishing between the two possible targets, initially one target would have very high contrast and the other very low. As observers correctly report which target has higher contrast, the contrast difference between the two alternative targets is progressively reduced. When this difference has become so small that observers start making errors in discriminating between the two alternative targets, it is slightly increased back. After a while, it is hence expected that the contrast difference will start oscillating around the point at which observers can just barely discriminate between the two alternative targets. By analyzing the data obtained through the course of such block of trials, we can compute, as shown below, the observer's threshold.

The specific staircase procedure employed consisted of the following:

1. After four correct answers in a row, the value of the parameter which distinguished between the two targets was decreased by a fixed amount;

2. After two incorrect answers, this parameter was increased by the same amount.

Typical staircases obtained in that manner are shown in **Fig. 6.2**.

6.2.3 Threshold Computation

Threshold was defined as the value of the stimulus parameter yielding 75% correct discrimination. Because the observer's psychometric function, i.e., the function which expresses the probability of correct answer for each possible value of the stimulus parameter, was only approximated by the trials collected during a few blocks, we used the following analytical method to reliably compute threshold values from the data.

The ideal psychometric function was modeled by a Weibull function with two degrees of freedom (Watson, 1990; Weibull, 1951). This function has a sigmoidal shape and is expressed as:

$$f(x) = 1 - \frac{1}{2} \exp\left(-(\alpha x)^{\beta}\right) \tag{6.1}$$

We fix β to the commonly used value of 2 and determine α using a linear least-squares method. In what follows, we define by $x_i, \{i \in [1..p]\}$ the values assumed by the stimulus parameter which is varied from trial to trial, by y_i the probability of correct answers for a given x_i (as measured from the observer's responses), and by n_i the total number of trials performed for that x_i.

We start by rewriting $f(x_i) = y_i$ as:

$$\alpha x_i = [-\ln(2 - 2y_i)]^{\frac{1}{\beta}} \tag{6.2}$$

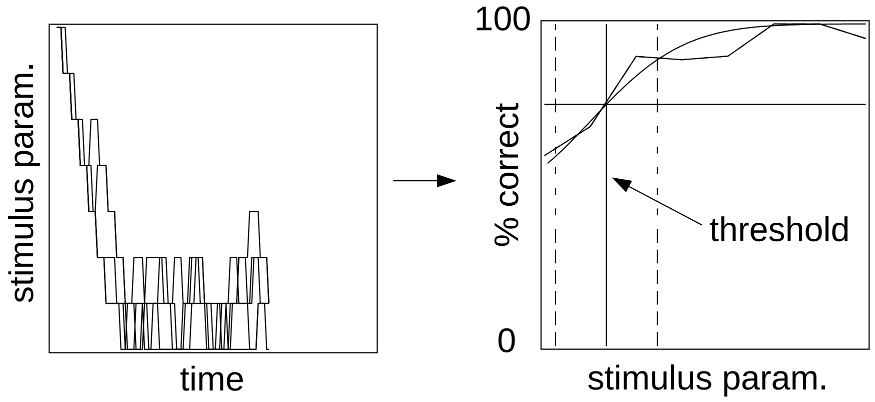


Figure 6.2: Staircase procedure and computation of threshold. The left plot shows the evolution with time of the stimulus parameter (e.g., contrast) which distinguishes between the two alternative targets, for 100 successive trials along the horizontal axis. After every four consecutive correct responses, the stimulus parameter was decreased by a fixed amount, and after every two consecutive incorrect responses, it was increased by the same amount. Several curves are shown for several blocks of trials with identical experimental settings. Although discrimination initially is easy, it progressively becomes more difficult as observers give correct answers. After approximately 30 trials, the value of the stimulus parameter starts oscillating around the observer's "threshold" value. The right plot represents the same data, but now plotted as a measure of probability of correct response for each stimulus parameter value (broken line). The smooth curve represents the fit of a Weibull function to the broken line curve. The "threshold" is computed from the intercept of this curve with the line of 75% correct performance. Error bars were computed using a binomial model for the data.

The least-squares minimization problem can then be written as:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \alpha = \begin{bmatrix} [-\ln(2-2y_1)]^{\frac{1}{\beta}} \\ [-\ln(2-2y_2)]^{\frac{1}{\beta}} \\ \vdots \\ [-\ln(2-2y_p)]^{\frac{1}{\beta}} \end{bmatrix} \tag{6.3}$$

For each value x_i, we consequently use the squared error:

$$e_i = \left[(-\ln(2-2y_i))^{\frac{1}{\beta}} - \alpha x_i\right]^2 \tag{6.4}$$

We wish to minimize, with respect to α, the total weighted error:

$$e = \sum_{i \in [1..p]} n_i^4 e_i \tag{6.5}$$

where the weighting of each datapoint by a coefficient n_i^4 is chosen empirically (we could have used n_i, various powers of n_i, or $\exp(n_i)$; the role of this weighting is to give low priority in fitting datapoints for which only very few trials were collected). At the minimum, $\partial e / \partial \alpha = 0$, which yields:

$$\sum_{i \in [1..p]} -2 n_i^4 x_i \left[(-\ln(2-2y_i))^{\frac{1}{\beta}} - \alpha x_i\right] = 0 \tag{6.6}$$

$$\alpha = \frac{\sum_{i \in [1..p]} n_i^4 x_i \left(-\ln(2-2y_i)\right)^{\frac{1}{\beta}}}{\sum_{i \in [1..p]} n_i^4 x_i^2} \tag{6.7}$$

Once α has been computed in this manner, threshold, x_T is obtained from the fitted Weibull function at the point where the probability of correct answer is $T = 0.75$ **(Fig. 6.2)**:

$$f(x_T) = T, \qquad \text{i.e.,} \qquad x_T = \frac{1}{\alpha}\left(-\ln(2-2y_i)\right)^{\frac{1}{\beta}} \tag{6.8}$$

To compute a confidence interval on this threshold x_T, we model our trials as being drawn from a binomial distribution. This gives us a confidence interval (standard deviation) for the value of T. We then use a linear approximation of the Weibull function f around threshold to convert that confidence on T into a confidence on x_T. Let's consider that a total of n trials are drawn from a binomial distribution with probability T; then the probability of obtaining m positive outcomes is:

$$P(m) = \binom{n}{m} T^m (1-T)^{n-m} \tag{6.9}$$

The mean of this distribution is nT and its variance $nT(1-T)$. This directly transcribes, for the proportion of correct trials m/n, into a mean of T and variance of $T(1-T)/n$. That is, by assuming that the observer's responses are drawn from such distributions, we can derive a variance on the threshold performance T. This requires us to estimate the total number of trials which were drawn for stimulus parameter x_T; since most of the time x_T does not exactly fall onto one of the discrete values x_i for which observer responses were collected, we estimate this total number using a linear interpolation between the total number of trials collected for the two closest values of x_i around x_T. We now simply convert the variance on T, $\sigma_T^2 = T(1-T)/n$, into a variance, $\sigma_{x_T}^2$ on x_T by using the slope of f at the point (x_T, T):

$$f'(x) = \frac{1}{2}\alpha^\beta x^{\beta-1} \exp\left(-(\alpha x)^\beta\right) \tag{6.10}$$

Hence, since $\sigma_{x_T} = \sigma_T / f'(x_T)$:

$$\sigma_{x_T} = \frac{\sqrt{T(1-T)/n}}{\frac{1}{2}\alpha^\beta x^{\beta-1} \exp\left(-(\alpha x)^\beta\right)} \tag{6.11}$$

From each set of staircase data corresponding to a given experimental setting, we are consequently in a position of computing the corresponding observer threshold, x_T, as well as an error bar on this threshold, σ_{x_T}.

6.2.4 Experimental Data Acquisition

Seven naive observers participated in the experiments. For each of our five tasks, observers received 15–24 blocks (of 100 trials each) of training and collected data for 45–90 blocks. Three observers completed all five tasks. Observers performed 10–12 blocks per daily hour-long session, typically shared between two different tasks. Each threshold estimate reflects 1,000–1,800 trials (two to three half-sessions). Four percent of all blocks were excluded because the staircase failed to converge. In total, data for in excess of 225,000 trials was collected. A set of programs was developed to automatically transfer the data to a master server after each session, create archives, and analyze results. This automatic process was of great value since it was necessary to analyze, for each observer, every day's data in order to guess suitable starting conditions for the following day's experiments.

As mentioned previously, targets appeared at a constant eccentricity of 4°, but at random polar angles. The targets were Gabor stimuli or superpositions of Gabor stimuli (spatial frequency 1.4 to 5.6cpd; half-width at half-maximum equal to the spatial period). At 4° eccentricity, cortical magnification is approximately 1.7mm/°, compared to approximately 10.0mm/° in the fovea (Cowey & Rolls, 1974). Thus, the stimulus diameter of approximately 0.7° (2 cycles at 2.8cpd) corresponds to approximately 1.2mm in primary visual cortex, that is, less than the average diameter of a hypercolumn (Blasdel & Salama, 1986). Presented in the fovea, the same stimulus would have excited approximately 35 hypercolumns.

Exp. 1 involved Gabor stimuli of different contrast (2.8cpd, vertical orientation, cosine phase). Observers reported which target had higher contrast and threshold was established in terms of incremental contrast (Nachmias & Sansbury, 1974; Legge & Foley, 1980; Foley, 1994). **Exp. 2** concerned Gabor stimuli of different orientation (2.8cpd, contrast 0.05–0.9, cosine phase). Observers reported which target was tilted clockwise from vertical, and threshold was measured in terms of tilt angle (Skottun *et al.*, 1987). **Exp. 3** used targets of different spatial frequency (vertical orientation, contrast 0.05–0.9, cosine phase). Observers reported which target had lower (coarser) spatial frequency, and threshold was measured in terms of the spatial frequency ratio in octaves (Skottun *et al.*, 1987). Two further experiments concerned the discrimination of two superimposed Gabor stimuli. One was present in both intervals ("mask," contrast 0.5, random phase) while the other appeared in only one interval ("target," 2.8cpd, vertical orientation, cosine phase). Observers reported which interval contained the target stimulus, and threshold was measured in terms of target contrast. In **Exp. 4**, mask orientation was varied between blocks (0–90°) but spatial frequency remained fixed (2.8cpd) (Phillips & Wilson, 1984). In **Exp. 5**, mask spatial frequency was varied between blocks (0.5–2oct.) but orientation was fixed (15° from vertical) (Wilson *et al.*, 1983).

6.3 Experimental Results

Psychophysical results from seven observers are shown in **(Fig. 6.3)**. On average, the detection threshold for Gabor patches of 2.8cpd, sub-tending approximately 0.7° at 4° eccentricity, is reached at a contrast of approximately 0.025 **(Fig. 6.3; Exp. 1)**. This is higher than the previously reported detection threshold contrast of 0.005, for gratings of 2.0cpd sub-tending 0.75° and presented in the fovea (Legge & Foley, 1980). Presumably, the discrepancy is due to the greater eccentricity (Greenlee, 1992; Thibos *et al.*, 1996) of our stimuli.

The minimal contrast increment threshold ("dipper") occurs at a pedestal contrast of $S_{fac} \approx 1.2$ times the detection threshold C_{th}, and is two to three times smaller than C_{th} **(Fig. 6.3; Exp. 1)**. This is in good agreement with previous studies, where the corresponding values are around 1.3 and 2.5, respectively (Legge & Foley, 1980). At higher

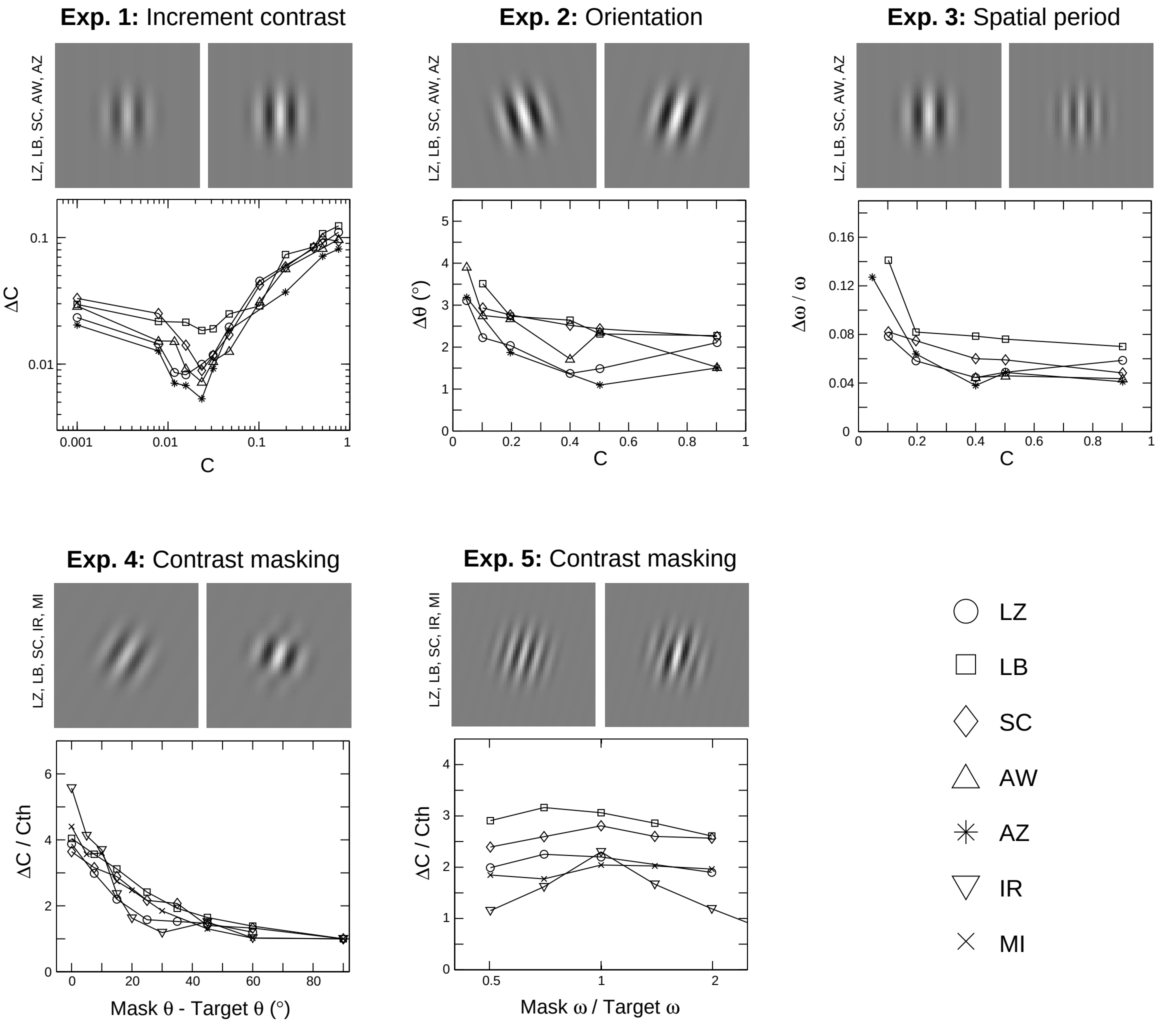


Figure 6.3: Experimental data from seven psychophysical observers. Three observers completed all five experiments (LB, LZ and SC), and others completed either two or three experiments (AW, AZ, IR, MI). All experiments involve a temporal two-alternative forced-choice discrimination between Gabor stimuli at 4° of eccentricity (insets). Exp 1: Contrast increment threshold, ΔC, as a function of contrast, C. Exps. 2 and 3: Orientation and relative spatial frequency discrimination thresholds, $\Delta\theta$ and $\Delta\omega/\omega$, as a function of contrast, C. Exp. 4: Contrast threshold elevation, $\Delta C/C_{th}$, as a function of relative mask orientation (C_{th} is the detection contrast threshold, leftmost point of Exp. 1). Exp. 5: Contrast threshold elevation, $\Delta C/C_{th}$, as a function of relative mask spatial frequency.

pedestal contrast, increment thresholds increase with a Weber exponent of approximately 0.6, consistent with the range of exponents (0.6 to 0.8) reported in previous studies (Legge & Foley, 1980; Wilson, 1980b; Foley, 1994). This close agreement is reassuring because it suggests that our peripheral thresholds reflect similar neural mechanisms than the foveal thresholds measured by previous authors.

In the limit of high contrast, average thresholds for orientation discrimination are approximately 2° **(Fig. 6.3; Exp. 2)** and average thresholds for spatial period discrimination approximately 0.05oct. **(Fig. 6.3; Exp. 3)**. Both values are at the upper end of the range reported previously for orientation discrimination (0.3–2°) and spatial period discrimination (0.02–0.05 oct) with relatively small patterns (Hirsch & Hylton, 1982; Burbeck & Regan, 1983; Wilson & Gelb, 1984; Vogels & Orban, 1990). Presumably, the relatively high discrimination thresholds are due to the eccentric stimulus location (Greenlee, 1992).

Our experiments on contrast masking are similar to those reported by Wilson and colleagues (Wilson *et al.*, 1983; Phillips & Wilson, 1984). In our case, the presence of a masking pattern of different orientation elevates contrast thresholds up to approximately 4-fold **(Fig. 6.3; Exp. 4)**, and a masking pattern of different spatial period elevates thresholds up to approximately 2.5-fold **(Fig. 6.3; Exp. 5)**. The corresponding threshold elevations reported previously for stimuli presented in the fovea are approximately 8-fold and 5-fold, respectively (Phillips & Wilson, 1984; Wilson *et al.*, 1983)). The difference may reflect either the smaller stimulus size or the greater stimulus eccentricity in our situation.

No threshold elevation is observed if target and masking patterns differ by more than 60° in orientation **(Fig. 6.3; Exp. 4)**. This contrasts with previous reports of an approximately 2.5-fold threshold elevation with orthogonal target and masking patterns ("cross-orientation inhibition" (Heeger, 1992; Foley, 1994)). However, the latter studies used large masker stimuli (extending over 7° × 5° of visual angle in the fovea (Foley, 1994)), and thus are likely to reflect interactions between both overlapping and non-overlapping visual filters. Our stimuli were designed to primarily probe interactions between overlapping visual filters.

Three of our experiments measured contrast increment threshold for patterns of 2.8cpd and a pedestal contrast of 0.5, allowing us to assess the consistency of our observations. In Exp. 1, target and mask patterns of identical orientation and phase produced an approximately 3.6-fold threshold elevation **(Fig. 6.3; Exp. 1)**. In Exp. 4, target and mask patterns varied in relative phase, but still produced an approximately 4-fold threshold elevation **(Fig. 6.3; Exp. 4)**. Finally, in Exp. 5, target and mask patterns differed in orientation (by 15°) and also varied in relative phase, producing an approximately 2.5-fold threshold elevation **(Fig. 6.3; Exp. 5)**.

6.4 Outlook

This dataset represents the first completely consistent dataset, to our knowledge, using five different tasks at 4° of eccentricity. As stated in the introduction of this Chapter, our two main motivations for acquiring such dataset were as follows.

First, we wanted to use localized stimuli in the near periphery, with the goal of only exciting a small number of cortical hypercolumns. Under these conditions, we believe that our data mostly reflects short-range, local cortical interactions, with little or no contribution from non-classical surround interactions. This is an important feature of this dataset in view of the modeling effort presented in the next Chapter, since only short-range interactions are included in our model. Indeed, we found that some of our results differed substantially from previously published foveal results; most importantly, the largest differences were found for the contrast masking experiments, which are believed to reflect cortical interactions among visual filters.

Second, we wished to obtain data on multiple tasks using a consistent set of stimuli and observers. Since previously published data concerned, to our knowledge, no more than two of our five experiments for any given observer, attempting to simultaneously model more than two experiments from such data would have created uncertainties and difficulties because of inter-observer variability. Indeed, we did observe significant inter-observer variability in our data **(Fig. 6.3)**. In addition, different experiments from different groups usually employed different experimental conditions, stimulus shapes, presentation paradigms, or presentation time, all of which are additional confounds when trying to fit a single model to all the data. The availability, through our effort, of the entire dataset of five experiments for three observers eliminates such uncertainties and difficulties. As a direct consequence, it becomes possible to address the question of whether a single, unified model can reproduce our dataset in its entirety. This question is the subject of the next Chapter.

Chapter 7 A Unifying Model of Spatial Vision

7.1 Overview

We propose a quantitative model relating the responses in small populations of visual filters to human psychophysics. The model consists of linear spatial filters which interact through non-linear excitatory and inhibitory pooling. Statistical estimation theory is used to derive human psychophysical thresholds from the population response. The model reproduces human thresholds for contrast increments, contrast masking, orientation discrimination and spatial period discrimination (as detailed in the previous Chapter) with a unique set of parameters. The success of this unified model suggests that a wide range of spatial vision thresholds reflect the same neural level of processing, presumably orientation- and spatial frequency-selective neurons in striate and extrastriate cortex.

The architecture of the model can be divided into three successive processing stages: The first stage consists of a set of linear visual filters tuned for spatial period and orientation; in the second stage, the outputs of the linear stage interact through non-linear excitatory and inhibitory connections. Physiological noise is introduced at the outputs of this stage. Finally, the third stage uses the entire population of noisy responses from the second stage to perform an ideal-observer discrimination between two stimulus patterns. The output of the model, for any two spatial patterns, is a quantitative measure of the system's performance at discriminating between these patterns. It can be directly compared to experimental data.

The present study is not primarily focused on possible detailed biophysical implementations of the model. Rather, the early stages of the model should be understood as a phenomenological, possibly consensual account for the observed physiology of visual neurons (or small populations of such neurons) in area V1. Furthermore, the last (decision) stage of the model, which is much more difficult to characterize in biological systems, is implemented in a more abstract manner using results from statistical estimation theory. In the discussion for this Chapter, we will however provide a number of cues towards possible biophysical implementations or circuits which may serve as plausible biological correlates for the model components.

In this Chapter, we detail the architecture of the model. We then present the methods developed to fit the model to our psychophysical data presented in the previous Chapter. Further, we present some of the methods which we devised to evaluate the quality and robustness of these fits. Because our model is only constrained by the data, using an automatic fitting procedure which involves no human bias, we then examine whether the best-fit model assumes neurally plausible components. We finally discuss the meaning of such successful account of a wide variety of spatial vision thresholds by a unique model, with a special emphasis onto which model features are critical for this success. This study will be extended in the next Chapter, in which we show that our model can also, in a very simple manner, account for the modulation of all of the measured psychophysical thresholds by focal visual attention.

Because this Chapter is mathematically much more involved than any other Chapter in the present thesis, we relay most of the mathematical details to a set of mathematical Appendices at the end of the Chapter.

7.2 The Model

The model comprises three successive levels: Linear filtering, divisive inhibition, and a statistically efficient decision strategy **(Fig. 7.1)**. We examine all of the model components in detail in this Section.

7.2.1 Linear Filters

We use a population of overlapping spatial filters, all centered at the same point of visual space, but tuned to a variety of spatial periods $\lambda \in \Lambda$ and orientations $\theta \in \Theta$. To facilitate comparison with cortical neurons and increase computational efficiency, filters are defined in terms of their tuning functions, rather than in terms of the spatial structure of their receptive field. For sinusoidal grating stimuli, we assume Gaussian tuning with respect to both the logarithm of spatial period λ_S (with standard deviation σ_λ) and orientation θ_S (with standard deviation σ_θ). Thus, the response of a filter with preferred period λ and preferred orientation θ to a sinusoidal grating of contrast C_S, period λ_S and orientation θ_S is given by:

$$E_{\lambda,\theta}(C_S, \lambda_S, \theta_S) = C_S A \exp\left(-\frac{(\log(\lambda_S) - \log(\lambda))^2}{2\sigma_\lambda^2} - \frac{(\theta_S - \theta)^2}{2\sigma_\theta^2}\right) \tag{7.1}$$

where A is a gain coefficient. Note that this definition specifies filters in the Fourier domain and disregards phase information. To obtain responses to arbitrary stimuli, two spatial filters with identical tuning but quadrature phase can be reconstructed and combined (Pollen & Ronner, 1981; DeAngelis *et al.*, 1992). The reconstructed spatial filters closely resemble Gabor functions and neuronal receptive fields **(Fig. 7.2)**.

Since we assume, for genericity reasons, that our filters have Gaussian tuning curves in response to sinusoidal grating stimuli, for the Gabor stimuli used in our experiments, we compute the responses numerically. Indeed, it is important to keep in mind that the tuning properties (i.e., the curve representing filter responses as one stimulus parameter is systematically varied) of any filter are dependent on the spectral characteristic of the stimulus **(Fig. 7.3)**; in what follows, the generic term of "tuning width" should be understood as the width of the tuning curve obtained for sinusoidal gratings.

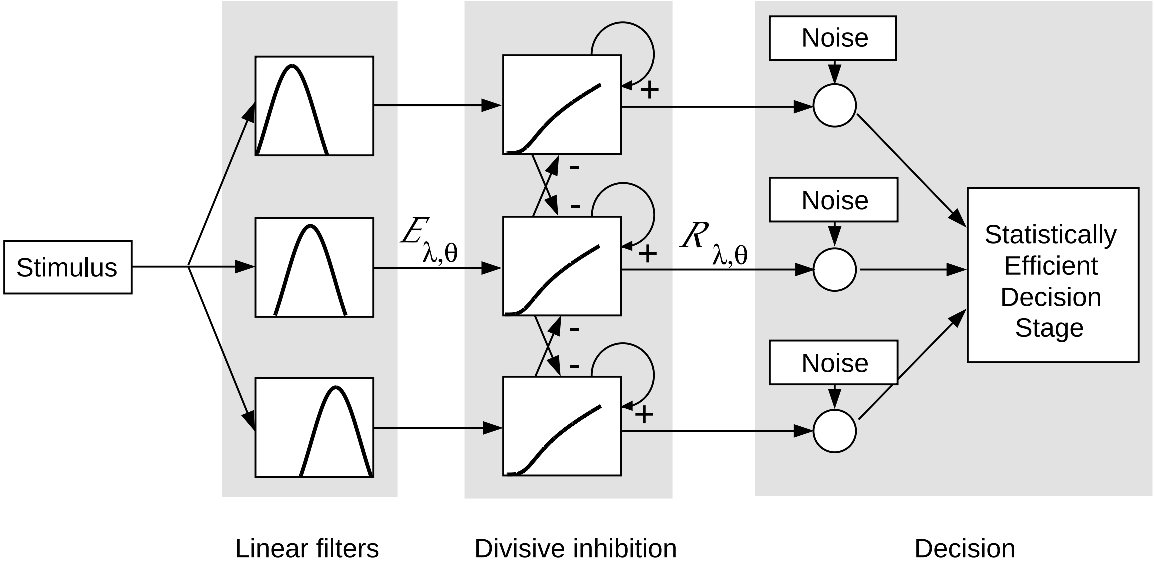


Figure 7.1: Model architecture, represented schematically in the style of Wilson and collaborators. The model consists of three successive stages: (1) A bank of linear visual filters tuned to different orientations and spatial frequencies, (2) non-linear interactions between visual filters in the form of a power law and divisive inhibition, (3) addition of independent noise and statistically efficient decision based on the entire filter population.

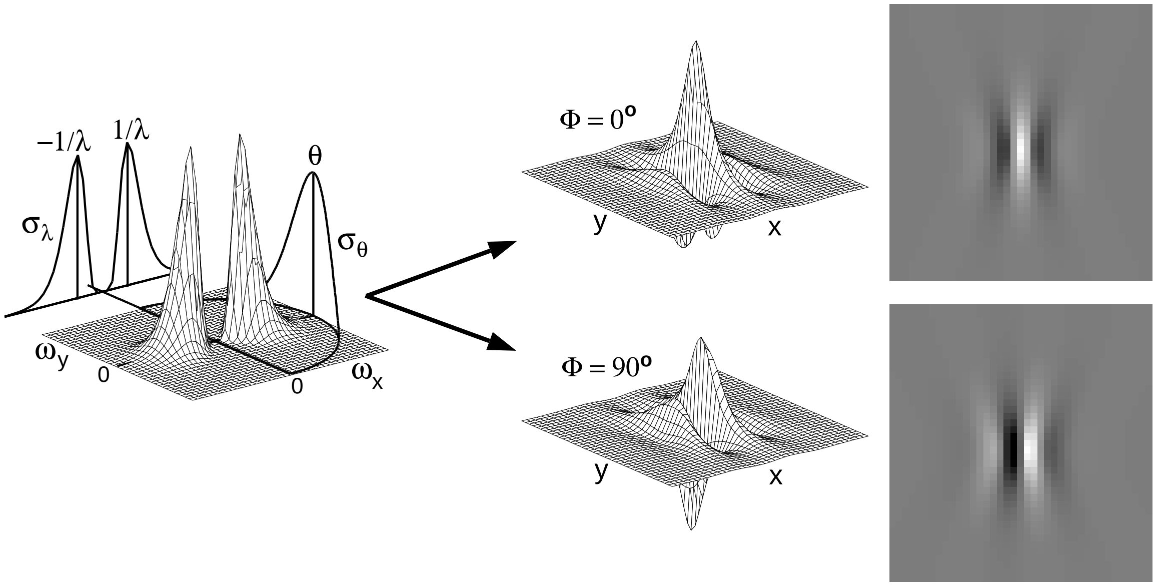


Figure 7.2: Visual filters used in the model. Filters are defined in terms of their separable Gaussian tuning functions for the orientation θ and logarithm of the spatial period λ of a sinusoidal grating stimulus (left). The spatial shape of the visual filter can be reconstructed though an inverse Fourier transform. Both even- and odd-symmetric filters are shown (middle and right). Their shape is very similar to the multi-lobed functions used by other models. Pairs of reconstructed even- and odd-symmetric filters can be used to compute the response to arbitrary stimuli.

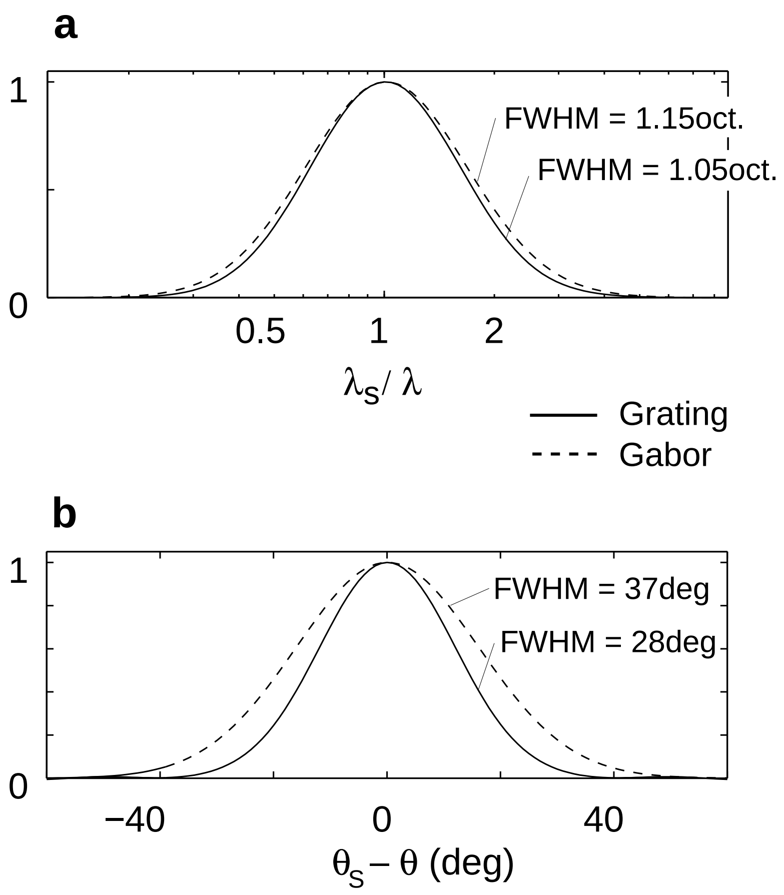


Figure 7.3: Tuning curves for grating and Gabor stimuli. Filters in the model are assumed to be specifically tuned to the spatial period λ and orientation θ of full-field sinusoidal grating stimuli. The Gaussian tuning curves for orientation **(a)** and spatial period **(b)** defined by Eq. 7.1 are shown here for gratings with period λ_S and orientation θ_S (solid curves). When Gabor stimuli are used, such as in our experiments, with same λ_S and θ_S, tuning curves for both spatial period and orientation are slightly broader (dashed curves).

7.2.2 Divisive Inhibition

The response of each linear filter is normalized relative to the total population response ("divisive inhibition" (Heeger, 1992)) (Wilson, 1993; Foley, 1994; Zenger & Sagi, 1996; Thomas & Olzak, 1997; Carandini *et al.*, 1998). The functional consequences of this normalization include a non-linear transducer (Nachmias & Sansbury, 1974; Wilson, 1980b) and reduced contrast-dependence of orientation and spatial frequency tuning (Wilson, 1993). The normalized response $R_{\lambda,\theta}$ of a filter tuned to (λ, θ) is:

$$R_{\lambda,\theta} = \frac{(E_{\lambda,\theta})^{\gamma}}{(S)^{\delta} + \sum_{(\lambda',\theta') \in \Lambda \times \Theta} W_{\lambda,\theta}(\lambda', \theta') \left(E_{\lambda',\theta'}\right)^{\delta}} + \eta \tag{7.2}$$

where

$$W_{\lambda,\theta}(\lambda', \theta') = \exp\left(-\frac{(\log(\lambda') - \log(\lambda))^2}{2\Sigma_{\lambda}^2} - \frac{(\theta' - \theta)^2}{2\Sigma_{\theta}^2}\right) \tag{7.3}$$

is a 2D Gaussian weighting function centered around (λ, θ) whose widths are determined by the scalars Σ_θ and Σ_λ **(Fig. 7.4)**. In **Eq. 7.2**, η is a positive constant representing background activity ("dark current"). The denominator includes a constant S and the weighted sum of all filter responses, and represents divisive inhibition. The exponents γ and δ determine the resulting transducer function. Depending on their values, one obtains a linear $(\gamma = 1, \delta = 0, S = 0)$, power-law $(S = 0)$, or sigmoidal $(S > 0, \gamma \geq \delta)$ transducer, with a saturating $(\gamma = \delta)$ or non-saturating $(\gamma \neq \delta)$ response.

It is convenient to replace A and S with two alternative parameters which are easier to interpret. We use the detection threshold C_{th} for a Grating stimulus of 2.8cpd and the position of the transducer's inflexion point S_{fac} $(S_{fac} = C_{\text{inflexion}}/C_{th})$. If all other model parameters are fixed, each choice of the pair (C_{th}, S_{fac}) yields a unique pair (A, S). The procedure used to compute (A, S) involves an iterative search over the value A which yields a detection contrast threshold of C_{th}; at each step during this search, S is computed using the closed-form solution detailed in Appendix 7.6.

7.2.3 Noise Model: Poisson$^{\alpha}$

Following normalization, independent Gaussian noise is added to each filter response. In analogy to visual cortical neurons, we assume that the variance increases with the response mean:

$$V_{\lambda,\theta}^2 = R_{\lambda,\theta}^{\alpha} \tag{7.4}$$

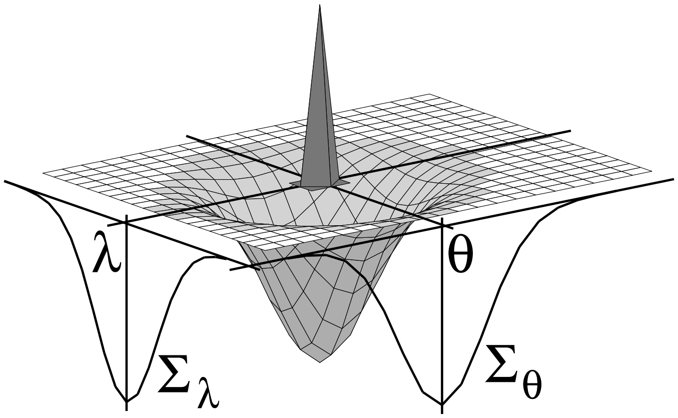


Figure 7.4: Schematic representation of the excitatory and inhibitory pool of units interacting with a given model unit. Each node in this diagram represents one unit in the population, tuned to orientation θ and spatial period λ. The pools are represented for the unit at the origin of the plot. The excitatory pool (top peak) is the single unit of interest. The inhibitory pool has a Gaussian shape in the orientation and spatial period domains, with widths determined by the scalar model parameters Σ_θ and Σ_λ.

where α is a constant. For visual cortical neurons, α is typically slightly larger than unity (Softky & Koch, 1993; Teich *et al.*, 1996; Geisler & Albrecht, 1997). Note that this noise assumption differs from most psychophysical models, which assume Gaussian noise of constant variance (Green & Swets, 1966; Sachs *et al.*, 1971; Wilson & Bergen, 1979; Legge & Foley, 1980; Thomas & Olzak, 1997).

7.2.4 Decision Stage

We use a statistically efficient decision stage to predict behavioral thresholds from the noisy responses of model units. This "ideal observer" decision accounts significantly better for our psychophysical data than the suboptimal (Pouget *et al.*, 1998) decision strategies used in other models (see Results).

We consider the noisy response vector $\mathcal{R} = \{R_{\lambda,\theta}; \lambda \in \Lambda, \theta \in \Theta\}$, with a mean given by **Eq. 7.2** and a variance by **Eq. 7.4**. For each stimulus attribute ζ, we postulate a statistic $T(\mathcal{R};\zeta)$ which estimates the value of ζ from the noisy response $\mathcal{R}$. The discrimination performance (see APPENDIX 7.8 and **Fig. 7.5**) in a 2AFC experiment involving two stimuli with attributes ζ_1 and ζ_2 is given by (Green & Swets, 1966):

$$\text{Performance} = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{|\text{mean}[T(\zeta_1)] - \text{mean}[T(\zeta_2)]|}{\sqrt{2(\text{var}[T(\zeta_1)] + \text{var}[T(\zeta_2)])}}\right) \tag{7.5}$$

where erf is the Normal error function. We assume that $T(\mathcal{R};\zeta)$ is "ideal" in the sense that it is both *unbiased* and *efficient*. The lack of bias implies that the estimator T exhibits no systematic bias towards either higher or lower values of ζ, in other words mean$[T(\zeta)] = \zeta$. Efficiency implies that var$[T(\zeta)]$ reaches the Cramér-Rao bound, the theoretical lower bound for the variance of any unbiased estimator (Scharf, 1991; Cover & Thomas, 1991). The Cramér-Rao bound equals the inverse of the Fisher information $\mathcal{J}(\zeta)$, such that var$[T(\zeta)] = 1/\mathcal{J}(\zeta)$. It follows that:

$$\text{Performance} = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{|\zeta_1 - \zeta_2|}{\sqrt{2(1/J(\zeta_1) + 1/J(\zeta_2))}}\right) \tag{7.6}$$

The Fisher information $J_{\lambda,\theta}$ for a single unit with response $R_{\lambda,\theta}$ and variance $V^2_{\lambda,\theta} = R^{\alpha}_{\lambda,\theta}$ is (APPENDIX 7.7):

$$J_{\lambda,\theta}(\zeta) = \left(\frac{\partial R_{\lambda,\theta}}{\partial \zeta}\right)^2 \left[\frac{1}{R^{\alpha}_{\lambda,\theta}} + \frac{\alpha^2}{2R^2_{\lambda,\theta}}\right] \tag{7.7}$$

In other words, information is distributed over the population and the units that respond maximally do not necessarily provide the most information. As the Fisher information is additive in the case of independent noise (Snippe & Koenderink, 1992; Cover & Thomas,

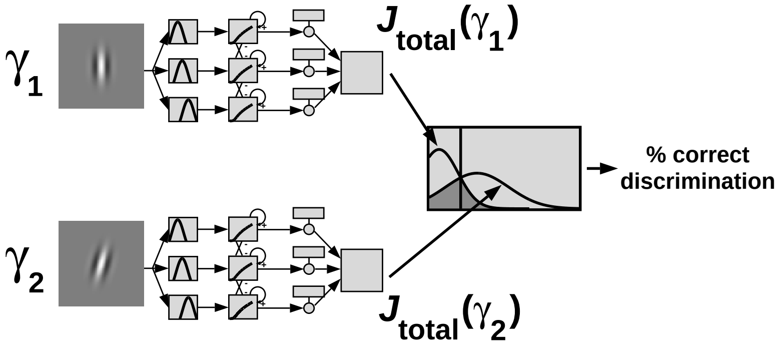


Figure 7.5: Ideal observer discrimination with our model. In a simulation of our psychophysical experiments, two alternative stimuli are presented to the model. The two stimuli differ only in their value of one of the stimulus parameters, γ (here, orientation). For each stimulus, our statistically efficient decision stage yields a Gaussian distribution whose mean is at the stimulus parameter γ and whose variance is the inverse of the Fisher information. Ideal observer discrimination then computes the probability of correct discrimination, based on the overlap between these two Gaussian distributions (Appendix 7.8).

1991), the total Fisher information for the entire population of units simply is:

$$\mathcal{J}(\zeta) = \sum_{\lambda,\theta} J_{\lambda,\theta}(\zeta) \tag{7.8}$$

Fig. 7.6 illustrates the distribution of Fisher information computed for stimulus contrast, orientation, and spatial frequency across units with different tuning properties. Note that information about each stimulus attribute is concentrated in different subpopulations of units. Due to our relatively sparse distribution of model units, we compute thresholds numerically (through iterative adjustments of ζ until threshold performance is reached). The closed-form solutions that hold for dense distributions of model units (Appendix 7.9 and (Mato & Sompolinsky, 1996; Seung & Sompolinsky, 1993)) are less accurate, especially near contrast threshold.

This decision stage can readily be generalized to other psychophysical paradigms. For example, performance in a "Yes/No" paradigm can be obtained by altering **Eq. 7.6**. A formal treatment of both the "Yes/No" and the "2AFC" paradigms in the context of ideal observer decision is presented in Appendix 7.8. Another possible generalization concerns decision uncertainty (Magnussen *et al.*, 1996; Dai *et al.*, 1996; Dai, 1994; Verghese & Stone, 1995). In the present experiments, there is no decision uncertainty, since always the same stimulus parameter (known to the observer) varies from trial to trial within each block. When this is not the case, and the decision involves several stimulus parameters, **Eq. 7.4** has to be generalized to $var[T] = \mathcal{J}(Z)^{-1}$, where Z is the vector of all relevant stimulus parameters.

7.2.5 Alternative Decision Stage

For comparison, we also used an alternative decision stage based on the Minkowski norm used in many popular psychophysical models ("Quick probability summation," (Quick, 1974; Wilson & Gelb, 1984; Phillips & Wilson, 1984; Foley, 1994; Bowne, 1990)). The discriminability of two stimuli with ζ_1 and ζ_2 is computed as:

$$\mathcal{D}(\zeta_1, \zeta_2) = \left[\sum_{(\lambda,\theta)\in\Lambda\times\Theta} |R_{\lambda,\theta}(\zeta_2) - R_{\lambda,\theta}(\zeta_1)|^Q \right]^{\frac{1}{Q}} \tag{7.9}$$

where Q is the Minkowski exponent (values above 3 yield similar results (Bowne, 1990)). Threshold is reached when $\mathcal{D} = 1$.

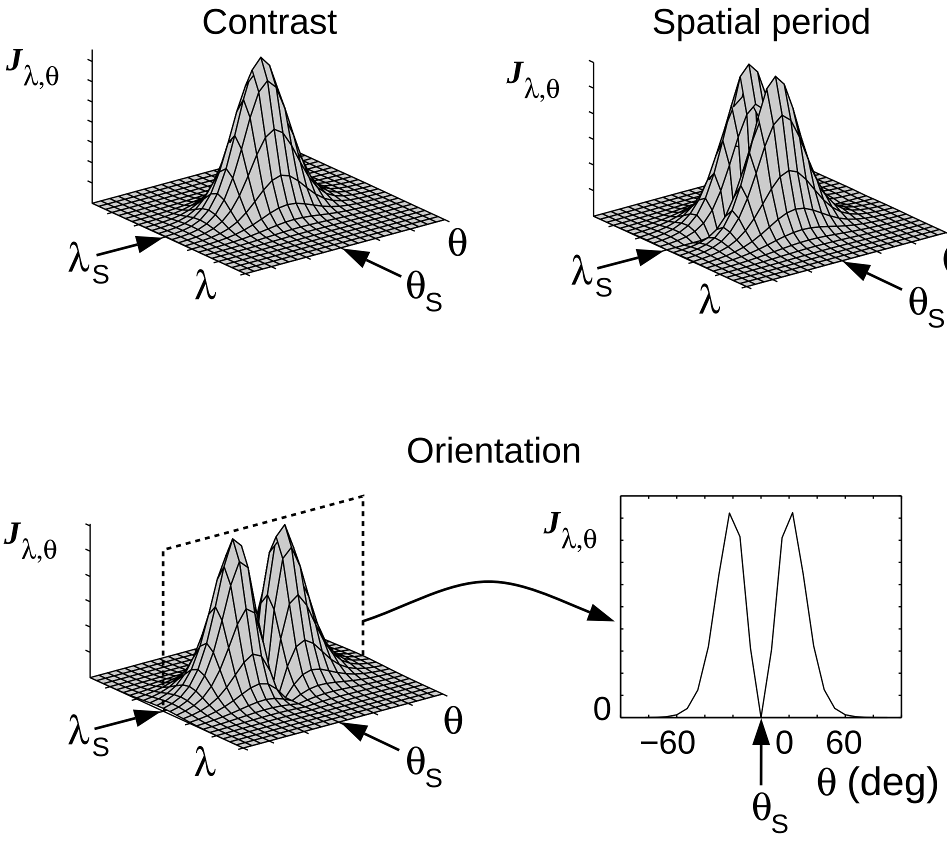


Figure 7.6: Fisher information with respect to stimulus orientation, contrast and spatial frequency. Fisher information is the inverse of the variance of an *unbiased and efficient* estimator of the stimulus parameter (Section 7.2.4 and Eq. 7.7). Each surface point represents the information encoded in the response of one model unit. The volume under the surface represents the total information encoded by a population of units with independent noise and tuned to 24 orientations and 24 spatial periods. Arrows indicate the spatial period λ_S and orientation $\theta_S = 0$ of the stimulus. Note that the unit tuned optimally for the stimulus does not contribute to the Fisher information for orientation or spatial period.

7.3 Model Predictions of Our Dataset

7.3.1 Predictions Using Our Standard Model

Specific Implementation of the Model Used

We implemented the model described above with 60 filters (spatial frequencies 1.4, 2, 2.8, 4 and 5.6cpd, and orientations 0, 15, 30, ..., 165°). The parameters for gain A, background activity η, orientation tuning width σ_θ, and spatial frequency tuning width σ_λ (in octaves) were identical for all filters, resulting in an overall total of only ten free parameters (**Table 7.1**). Best fitting parameter values were computed separately for each of the three subjects who completed all experiments (LZ, LB, SC). The total fit error was computed as the root of the squared percentage errors summed over all data points. Using percentage errors rather than absolute errors ensured that all data points carried numerically equal weight.

Fitting Procedure

Best fits were computed with two automatic procedures, with no operator bias towards "plausible" values. The first procedure used a ten-dimensional simplex algorithm (Press *et al.*, 1992) with simulated annealing overhead (Metropolis *et al.*, 1953; Kirkpatrick *et al.*, 1983). From several randomly chosen starting points, the annealing schedule was initiated with a "temperature" t that induced random parameter variations of ±15% at each transformation of the simplex. Every 200 simplex transformations, the annealing amplitude was reduced on a schedule $t \propto \log(1+k)$, where k is the number of simplex iterations. The process was terminated when the annealing amplitude became smaller than ±0.25%, at which point a final deterministic fit was carried out. The annealing schedule was sufficiently slow to ensure eventual convergence towards the global minimum (Geman & Geman, 1984). The second procedure also used several randomly chosen starting points but approached the best fit with the help of Powell's deterministic algorithm using Brent's minimization method (Press *et al.*, 1992).

Because of the high computational demands of our minimization problem, several machines were used to evaluate diverse variations of the model, diverse observers, or diverse parameter starting points **(Fig. 7.8)**. A WWW-based interface was developed to dispatch and collect results to and from these remote machines, as well as to monitor their health in real-time **(Figs. 7.8 and 7.9)**. This technical development proved critical to the management of the high complexity of our study. Indeed, a complete fit took several days to converge; with these tools, we were able to properly use in excess of 15 simultaneous machines without introducing any uncertainty in the collection of results and their interpretation.

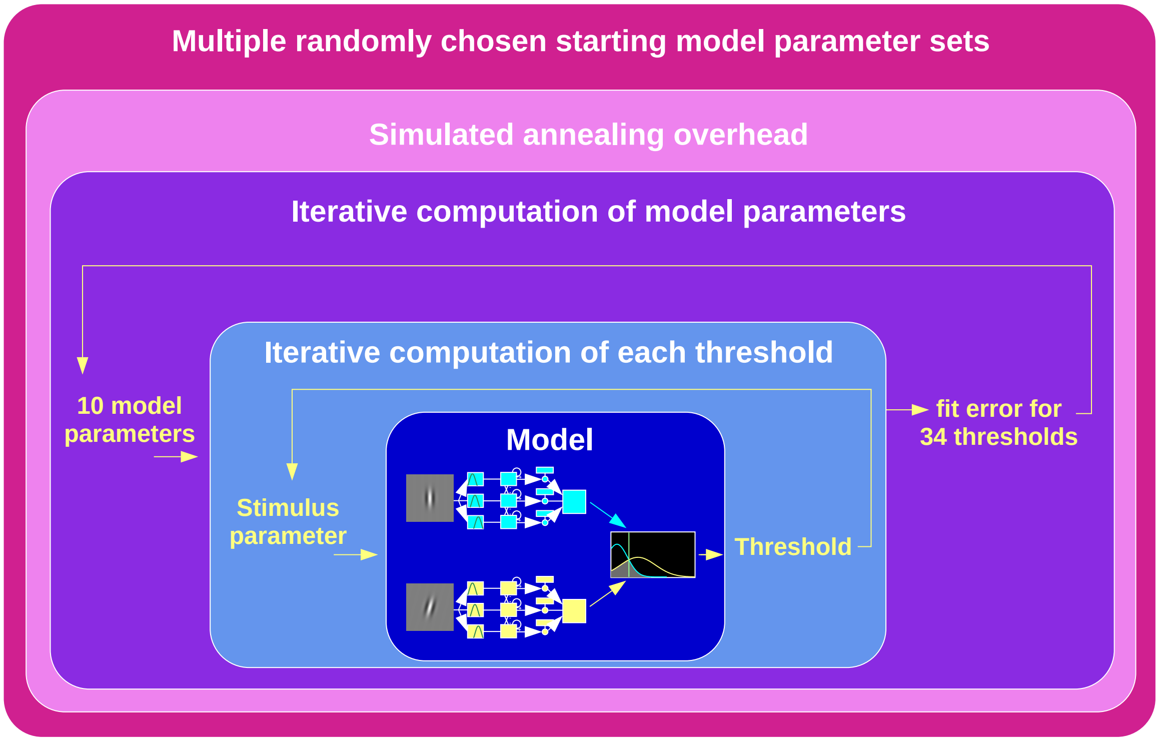
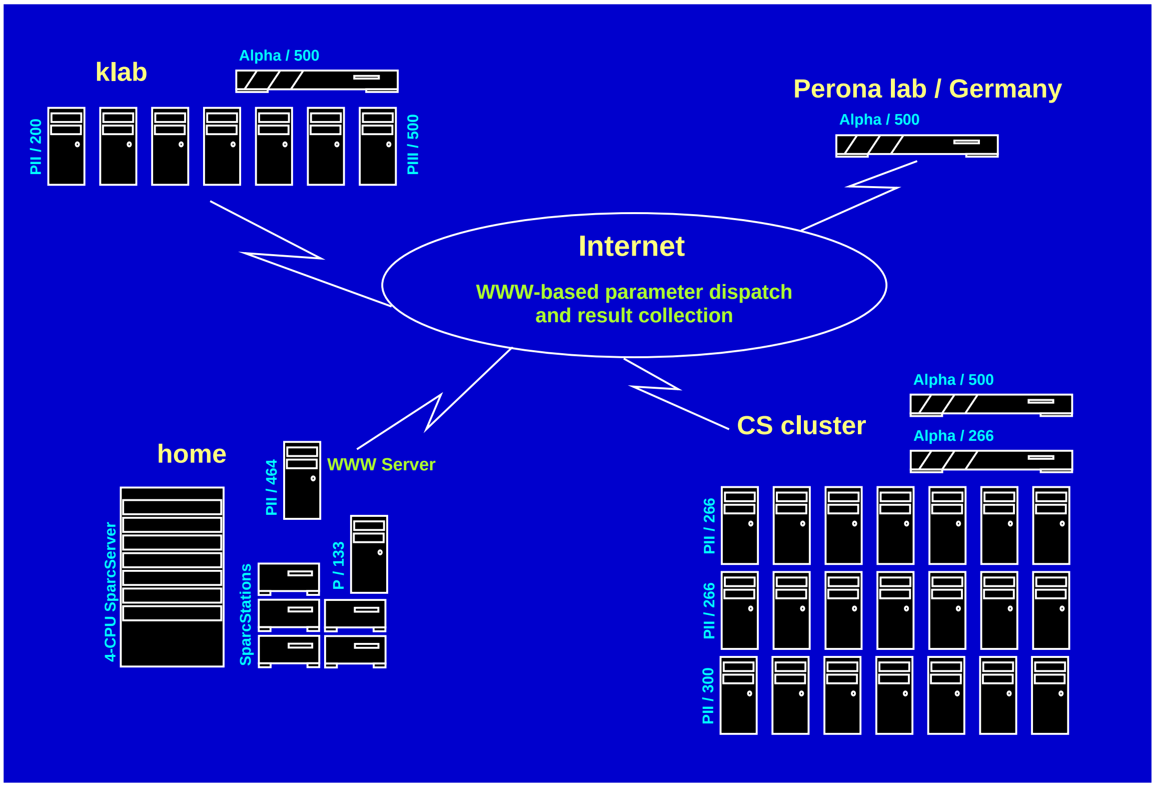
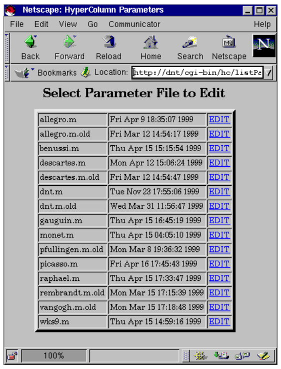

Multiple randomly chosen starting model parameter sets
Simulated annealing overhead
Iterative computation of model parameters
Iterative computation of each threshold
10 model parameters
fit error for 34 thresholds
Model
Stimulus parameter
Threshold

Figure 7.7: Procedure used to fit the model to the data. Each threshold was computed with the model as described in this part, using an iterative method which adjusted the stimulus parameter differentiating between the two possible targets until threshold performance was used. Because we use a general noise model, sparse filter population, and non-linear filter interactions, such computation could not be carried out in closed-form. After all thresholds were predicted by the model, a global fitting error was computed. A ten-dimensional downhill simplex was used to adjust the ten model parameters such as to minimize the fitting error. A simulated annealing overhead was added to ensure convergence towards the global minimum of the error. Finally, the algorithm was allowed to run from multiple randomly chosen starting points, and we verified that it always converged towards the same minimum error. For supplemental verification purposes, a second minimization method was implemented, the deterministic multidimensional optimization method of Powell, used as an overhead to a 1D Golden-section bracketing algorithm followed by Brent's 1D minimization (Press *et al.*, 1992).



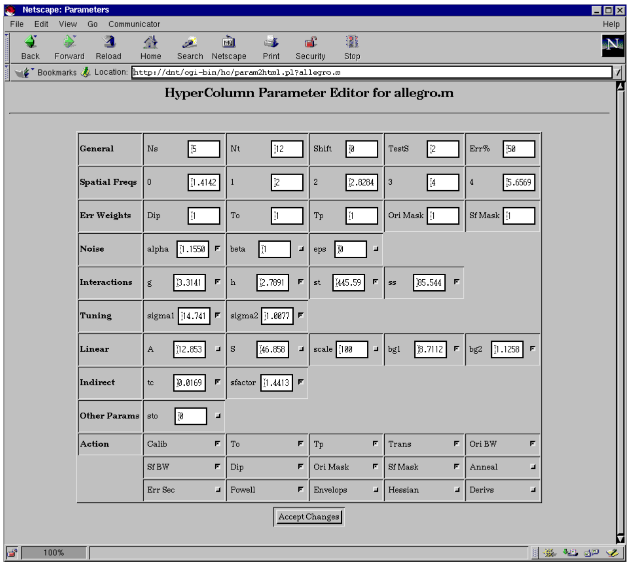
WWW-based control:


Netscape: HyperColumn Parameters
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape
Bookmarks Location:

http://dnt/cgi-bin/hc/listPa

Select Parameter File to Edit

allegro.m	Fri Apr 9 18:35:07 1999	EDIT
allegro.m.old	Fri Mar 12 14:54:17 1999	EDIT
benussi.m	Thu Apr 15 15:15:54 1999	EDIT
descartes.m	Mon Apr 12 15:06:24 1999	EDIT
descartes.m.old	Fri Mar 12 14:54:47 1999	EDIT
dnt.m	Tue Nov 23 17:55:06 1999	EDIT
dnt.m.old	Wed Mar 31 11:56:47 1999	EDIT
gauguin.m	Thu Apr 15 16:45:19 1999	EDIT
monet.m	Thu Apr 15 04:05:10 1999	EDIT
pfullingen.m.old	Mon Mar 8 19:36:32 1999	EDIT
picasso.m	Fri Apr 16 17:45:43 1999	EDIT
raphael.m	Thu Apr 15 17:33:47 1999	EDIT
rembrandt.m.old	Mon Mar 15 17:15:39 1999	EDIT
vangogh.m.old	Mon Mar 15 17:18:48 1999	EDIT
wks9.m	Thu Apr 15 14:59:16 1999	EDIT

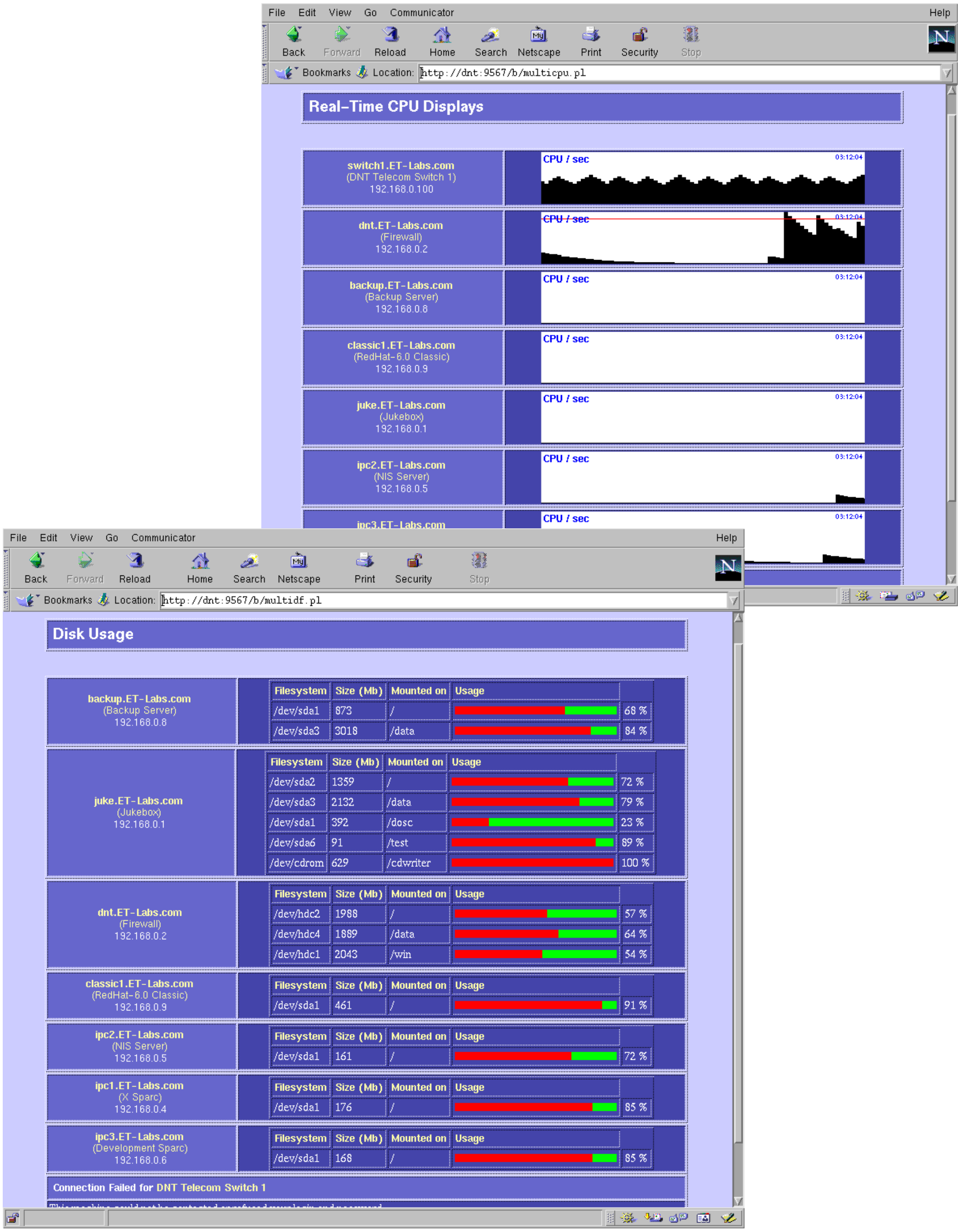

Netscape: Parameters
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location:

http://dnt/cgi-bin/hc/param2html.pl?allegro.m

HyperColumn Parameter Editor for allegro.m

General	Ns 5	Nt 12	Shift 0	TestS 2	Err% 50
Spatial Freqs	0 1.4142	1 2	2 2.8284	3 4	4 5.6569
Err Weights	Dip 1	To 1	Tp 1	Ori Mask 1	Sf Mask 1
Noise	alpha 1.1550 ☑	beta 1 □	eps 0 □		
Interactions	g 3.3141 ☑	h 2.7891 ☑	st 445.59 ☑	ss 85.544 ☑	
Tuning	sigma1 14.741 ☑	sigma2 1.0077 ☑			
Linear	A 12.853 □	S 46.858 □	scale 100 □	bg1 8.7112 ☑	bg2 1.1258 ☑
Indirect	tc 0.0169 ☑	sfactor 1.4413 ☑			
Other Params	sto 0 □				
Action	Calib ☑	To ☑	Tp ☑	Trans ☑	Ori BW ☑
	Sf BW ☑	Dip ☑	Ori Mask ☑	Sf Mask ☑	Anneal □
	Err Sec □	Powell ☑	Envelops □	Hessian □	Derivs □

Figure 7.8: Multi-machine system used to carry out the model fits. Several minimizations were dispatched in parallel to multiple machines in our and other laboratories. In order to properly manage such parallel processing, a WWW-based interface was created, which allowed each machine to automatically report on its results to a master server, and receive new instructions for a further fit.


File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location:
CPU / sec
03:12:04

http://dnt:9567/b/multicpu.pl

Real–Time CPU Displays

switch1.ET-Labs.com
(DNT Telecom Switch 1)
192.168.0.100

dnt.ET-Labs.com
(Firewall)
192.168.0.2

backup.ET-Labs.com
(Backup Server)
192.168.0.8

classic1.ET-Labs.com
(RedHat-6.0 Classic)
192.168.0.9

juke.ET-Labs.com
(Jukebox)
192.168.0.1

ipc2.ET-Labs.com
(NIS Server)
192.168.0.5

ipc3.ET-Labs.com

http://dnt:9567/b/multidf.pl

Disk Usage

backup.ET-Labs.com
(Backup Server)
192.168.0.8

Filesystem	Size (Mb)	Mounted on	Usage	
/dev/sda1	873	/		68 %
/dev/sda3	3018	/data		84 %

juke.ET-Labs.com
(Jukebox)
192.168.0.1

Filesystem	Size (Mb)	Mounted on	Usage	
/dev/sda2	1359	/		72 %
/dev/sda3	2132	/data		79 %
/dev/sda1	392	/dosc		23 %
/dev/sda6	91	/test		89 %
/dev/cdrom	629	/cdwriter		100 %

dnt.ET-Labs.com
(Firewall)
192.168.0.2

Filesystem	Size (Mb)	Mounted on	Usage	
/dev/hdc2	1988	/		57 %
/dev/hdc4	1889	/data		64 %
/dev/hdc1	2043	/win		54 %

classic1.ET-Labs.com
(RedHat-6.0 Classic)
192.168.0.9

Filesystem	Size (Mb)	Mounted on	Usage	
/dev/sda1	461	/		91 %

ipc2.ET-Labs.com
(NIS Server)
192.168.0.5

Filesystem	Size (Mb)	Mounted on	Usage	
/dev/sda1	161	/		72 %

ipc1.ET-Labs.com
(X Sparc)
192.168.0.4

Filesystem	Size (Mb)	Mounted on	Usage	
/dev/sda1	176	/		85 %

ipc3.ET-Labs.com
(Development Sparc)
192.168.0.6

Filesystem	Size (Mb)	Mounted on	Usage	
/dev/sda1	168	/		85 %

Connection Failed for DNT Telecom Switch 1

Figure 7.9: Multi-machine monitoring system. Because of the distant physical location of several machines used for our fits, and of the tendency of their interactive users to crash or reboot them, a WWW-based remote Unix administration system was used to monitor their health. This system was a prototype of a broadcast-capable Unix system and network administration tool which I was by that time developing with Distributed Network Technology, Inc. (Denver, CO) and Kaptech, S.A. (Paris, France).

Fit Results

For each of the three complete data sets, the best fit of the model is shown in **Fig. 7.10**. The model accounts quantitatively for all observations and in almost all cases predicted and measured thresholds agree to within the accuracy of the measurement. The model reproduced the dipper-shaped results of Exp. 1, the almost "flat" contrast dependence in Exps. 2 and 3, as well as the masking effects of Exps. 4 and 5. Some of the reasons for this success become clear when we consider the effective properties of individual model units. The effective transducer function was sigmoidal at low contrast and non-saturating at high contrast (**Fig. 7.11a**), explaining the results of Exp. 1 and their conformance to Guilford's law, which states that $\Delta C \propto C^x, 0.5 \leq x \leq 1$. The effective tuning for orientation was about 30% sharper than that of the linear units prior to the interactions **(Fig. 7.1)**, and the effective tuning for spatial period was about 35% narrower (**Fig. 7.11cd**), explaining the low thresholds for discriminating orientation and spatial period. The flat contrast dependence of these thresholds will be considered in a separate section.

The parameter values which yielded the best fit for each of the three complete data sets are listed in **Table 7.1**. Each value is given with a "tolerance range" (expressed as a percentage). This is the range in which a given parameter can vary such that the total fit error remains within 5% of its minimal value, when all other parameters are optimized to keep the fit error as small as possible (Appendix 7.10). Most tightly constrained (to within 2% of their respective values) were the exponents γ and δ. In the observer average, the optimal values were approximately $\gamma = 3.5$ and $\delta = 3.0$, that is, substantially larger than the values inferred from physiology (Heeger, 1992; Carandini & Heeger, 1994). The tuning widths σ_θ, σ_λ, noise exponent α, contrast threshold C_{th}, and transducer inflexion S_{fac} were constrained to within 15% of their respective values. The width of orientation tuning was approximately 35° full-width at half-maximum (FWHM) at the linear and approximately 25° at the non-linear stage. Similarly, the width of spatial period tuning was approximately 1.5oct and 0.8oct FWHM at the linear and non-linear stages, respectively. The noise exponent α was slightly larger than unity (values near 1.1). Less well constrained were the values for orientation pooling width Σ_θ (25° FWHM to within 25%) and for background activity η (to within 40%). The spatial period pooling width Σ_λ was only weakly constrained by our data, mostly because the measured contrast masking thresholds depended only weakly on the spatial period of the mask.

Stability of the Fits

Envelopes of Model Predictions. As the computation of tolerance values is based on the entire data set, it does not reveal the relative importance of different *parts* of the data set. To obtain some information on this point, we computed threshold predictions for the

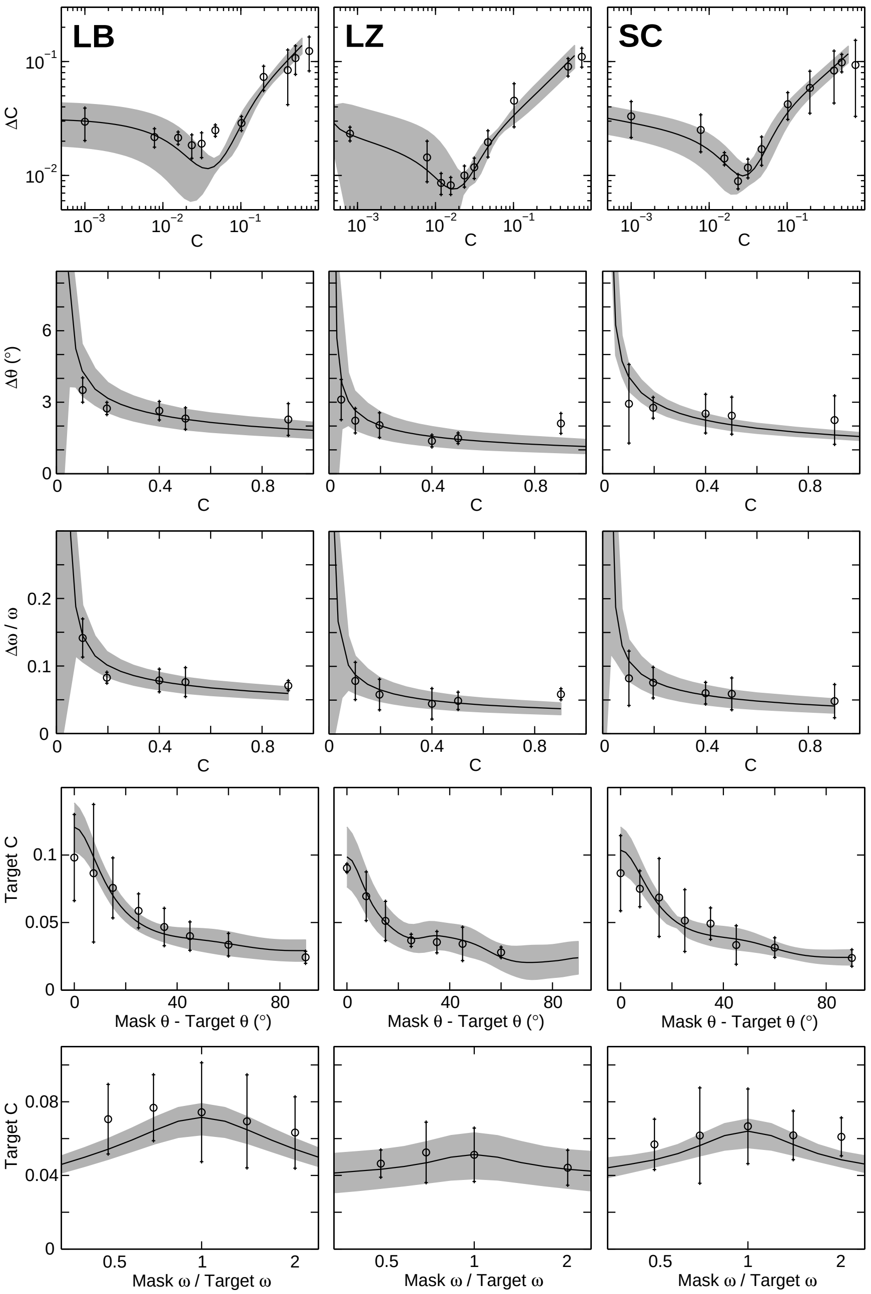


Figure 7.10: Measured and predicted thresholds for three observers (LB, LZ and SC). Measured thresholds (symbols with error bars) represent the mean and standard deviation for each observer. Predicted thresholds represent the optimal model fit (solid line) and the family of all model fits with up to 5% higher fit error (grey regions). Predicted thresholds are close to measure thresholds, often to within the accuracy of the measurement. In general, the grey regions closely "hug" the measured thresholds, demonstrating that the model fit is robust and not accidental. The narrow parts of the grey regions indicate which parts of the data constitute particularly tight constrains for the model.

Name	Symbol	LB	LZ	SC
Excitatory exponent	γ	$3.2 \pm 0.02\%$	$3.8 \pm 1\%$	$3.6 \pm 0.1\%$
Inhibitory exponent	δ	$2.7 \pm 1\%$	$3.0 \pm 2\%$	$3.0 \pm 1\%$
Noise exponent	α	$1.1 \pm 8\%$	$1.3 \pm 2\%$	$1.0 \pm 1\%$
Background activity	η	$3.8 \pm 31\%$	$1.7 \pm 10\%$	$12.3 \pm 40\%$
Spatial period tuning width	σ_λ (oct)	$0.76 \pm 9\%$	$0.54 \pm 10\%$	$0.64 \pm 9\%$
Orientation tuning width	σ_θ (°)	$17.8 \pm 16\%$	$12.5 \pm 7\%$	$16.5 \pm 6\%$
Spatial period pooling width	Σ_λ (oct)	$7.1 \pm 297\%$	$5.7 \pm 550\%$	$1.2 \pm 291\%$
Orientation pooling width	Σ_θ (°)	$14.5 \pm 23\%$	$12.0 \pm 16\%$	$11.0 \pm 12\%$
Contrast detection threshold	C_{th}	$0.026 \pm 13\%$	$0.025 \pm 3\%$	$0.026 \pm 13\%$
Transducer inflexion point	S_{fac}	$1.26 \pm 1\%$	$0.62 \pm 8\%$	$0.80 \pm 8\%$
Residual fit error		16%	14%	17%
Linear spatial period FWHM*	(oct)	1.79	1.27	1.51
Linear orientation FWHM	(°)	41.9	29.4	39.1
Pooled spatial period FWHM	(oct)	1.12	0.68	0.71
Pooled orientation FWHM	(°)	27.5	18.0	26.1

* Full-width at half-maximum (FWHM) is computed as $2\sigma\sqrt{2\log(2)}$ for the linear filters, and is measured at the output of the second stage (for a grating of contrast 0.1) for the pooled filters.

Table 7.1: Best-fit model parameters for observers LZ, LB and SC. Note that $\alpha = 1$ corresponds to Poisson noise. $\gamma - \delta$, always positive for our data, determines the asymptotic slope of the contrast response function.

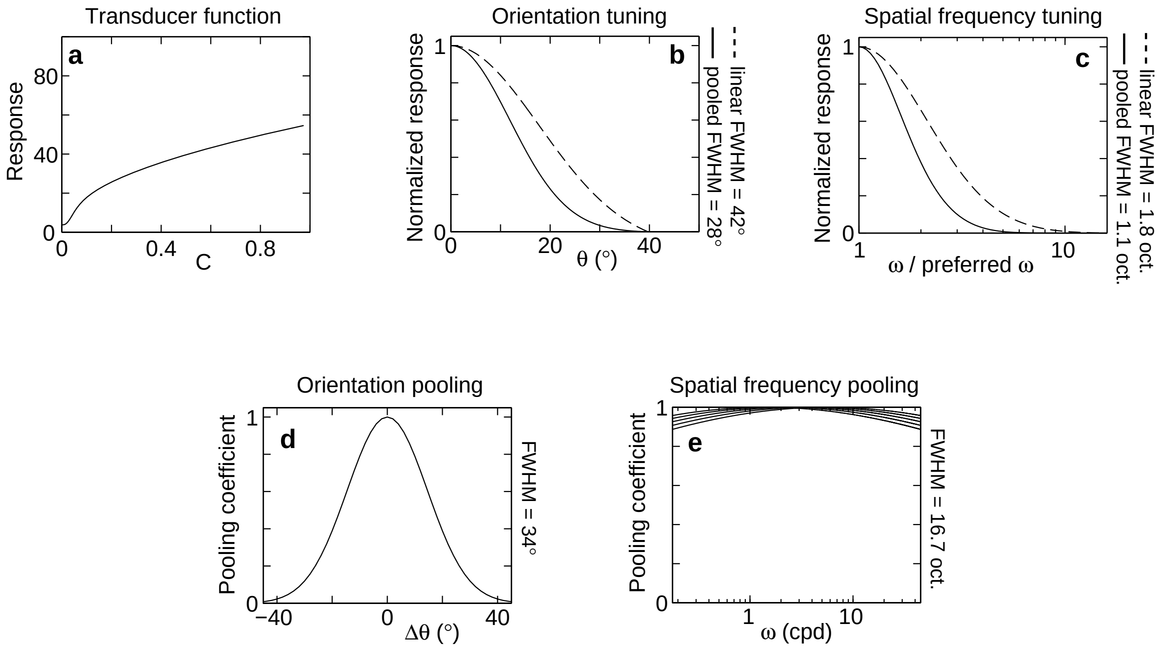


Figure 7.11: Functional properties of the optimal model for observer SC. **(a)** The effective contrast response function exhibits the sigmoidal shape postulated by most psychophysical models. **(b)** and **(c)** The effective tuning functions for orientation and spatial frequency are approximately Gaussian, but 30-40% narrower than the original tuning functions (dashed lines). **(d)** and **(e)** Relative weights with which different filters contribute to divisive inhibition. Inhibition derives from filters tuned to similar orientations (difference less than approximately 40°). The range of spatial frequencies contributing to divisive inhibition is only poorly constrained by the data.

family of all models yielding fit errors less than 5% above the optimal fit (Appendix 7.10). The envelope of the resulting predictions is narrow in the "more important" parts and broad in the "less important" parts of the data set (**Fig. 7.10**). By this criterion, the most important parts of the data set are the high-contrast regimes of Exps. 1, 2 and 3 ($C > 0.1$). Exps. 4 and 5 appeared important as well, as the predicted envelopes are quite narrow. In the low-contrast regime of Exp. 1, the predicted envelopes are wide but only concern 3% of the entire contrast range of that experiment. Least important are the low-contrast regimes of Exps. 2 and 3, where predictions diverge widely. This analysis shows that the data did closely constrain the model and that, overall, the model was not sensitive to small departures from the optimal parameter values. This demonstrates that the optimal fit is robust and non-accidental.

Cross-Sections of the Error-of-Fit Function. Another important test was to evaluate the precision with which each model parameter was constrained by the entire experimental dataset. This was investigated by computing local orthogonal cross-sections of the fitting error along individual parameters **(Fig. 7.12)**. Although these do not capture all possible local parameter variations around the best-fit point, they are computationally tractable and provide a reasonable idea of how sensitive the model's predictions are with respect to each of the parameters. This test also is important because our computation or tolerance ranges on the best-fit parameters involved a second-order approximation of the fitting error by its Hessian at the best-fit point (Appendix 7.10). The explicitly calculated cross-sections prove us that such approximation was valid (since we can see in **Fig. 7.12** that the error function is well behaved and can reasonably be approximated by a hyperparaboloid). The results indicated that all parameters were – to varying degrees – constrained by the data, and were in good agreement with the results obtained with our analytical method when all parameters are allowed to vary (Appendix 7.10).

Partial datasets. Although different experiments to some extent constrain different properties of the model, there is no straightforward correspondence between each experiment and the model parameters it constrains. Attempts to fit a randomly initialized model to partial datasets (e.g., including only one experiment) did not yield readily interpretable results, because each subset of the entire dataset inevitably left some parameters unconstrained. However, it is informative to investigate how the local error sections of **Fig. 7.12** would vary if they were computed only for partial datasets. We investigated all 31 possible binary combinations of the 5 experiments (obtained by assigning a weight of either 0 or 1 to each experiment's contribution to the fitting error). Two characteristic changes in the error sections were expected when computing the fitting error from a given partial combination: First, a shift of the minimum in a given section would indicate that the five experiments did not agree on the best value for the corresponding parameter. The best fit found would then represent a compromise between conflicting constraints imposed by the various experi-

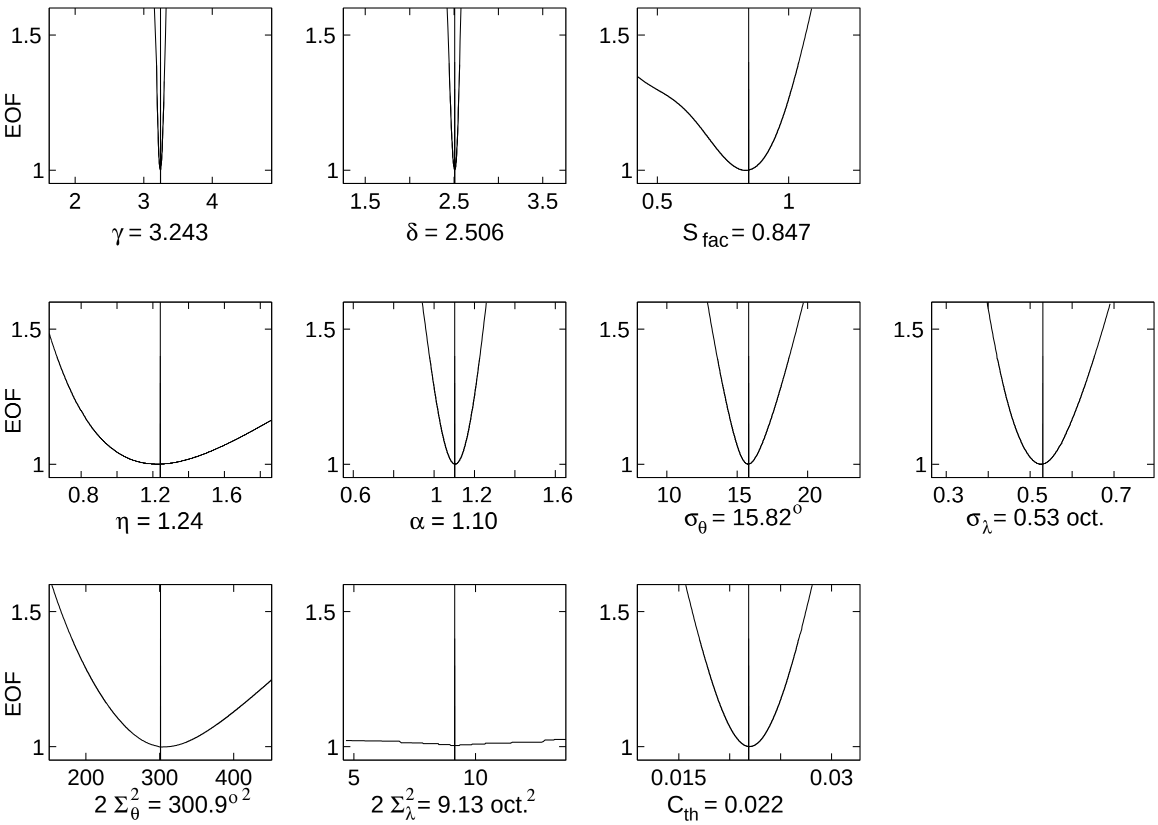


Figure 7.12: Local variations of the error of fit (EOF) around the best fit point, for an average (not shown) of the datasets of observers LB, LZ and SC. Each of the ten parameters was individually varied by up to ±50% around its best-fit value, in steps of 0.25%. Those parameters for which steeper curves around the best-fit point are observed, such as the excitatory (γ) and inhibitory (δ) exponents, are more strongly constrained by the data. The spatial period pooling width (Σ_λ) was only very weakly constrained. This method is less complete than the formal analysis presented in Appendix 7.10 and used to derive parameter tolerance ranges in **Tables 7.1** and **7.2**. Indeed, here, all parameters except the one varied in a given section are held fixed; in contrast, in the method of Appendix 7.10, all parameters are allowed to arbitrarily vary (within the only constraint that the fitting error should not exceed a given threshold value). However, this figure demonstrates that the error function is well behaved around the best-fit point and can reasonably be approximated by a hyperparaboloid, which is a prerequisite for the method of Appendix 7.10 to be valid.

ments, rather than an optimum satisfying all experiments simultaneously. Second, study of the local curvature of the error function around the minimum in each section would indicate which combination of experiments – or which individual experiment – is mainly responsible for constraining a given parameter. This would allow us to establish a correspondence between experiments and model parameters.

Of the 31 possible binary combinations of the five experiments, the five in which one experiment is omitted from the computation of the fitting error proved the most interesting: Each of these combinations reveals, for each parameter, whether the omitted experiment is important for constraining that parameter. For instance, if σ_θ became poorly constrained when the orientation discrimination experiment (Exp. 2) was ignored, we would conclude that Exp. 2 plays a critical role in constraining σ_θ. For each partial dataset, every parameter was varied around its best-fit value, and the partial fitting error was computed. The local minimum in each of the resulting error sections indicated the locally closest parameter value which fitted best the partial dataset (see ranges in **Table 1**). In addition, we computed a measure of how strongly each parameter was constrained by each partial dataset **(Fig. 7.13)**.

Remarkably, ignoring any one experiment did not change any parameter dramatically, as is detailed below and in **Table 1**. Exp. 1 was mainly responsible for constraining S_{fac} and C_{th}, although ignoring any one experiment yielded local minima for both parameters in $\pm 2\%$ range (*i.e.*, all experiments agreed very closely on these parameter values). Exp. 2 constrained the orientation tuning width (σ_θ), to a value 8% lower than would have satisfied the other experiments. Exp. 3 constrained the spatial period tuning width (σ_λ), to a value within $\pm 3\%$ of that satisfying the other experiments. Exp. 4 constrained the orientation pool width (Σ_θ), and was in part responsible for keeping the noise level (α) low. Exp. 5 did not appear to significantly constrain any parameter, while we would have expected it to constrain Σ_λ, if the measured data had been more reliable and more complete. Exp. 5 was, however, in part responsible for keeping the noise level (α) high. It cannot be argued that Exps. 4 and 5 conflicted with respect to the parameter α, since predicted values differed by less than 3%. The pooling exponents γ and δ (*i.e.*, the high-contrast regime of the transducer function) were the most strongly constrained parameters, and were constrained by all partial datasets to values in $\pm 1\%$ range. Finally the background activities (ϵ, η) and the spatial period pooling width (Σ_λ) were poorly constrained, but ignoring any one experiment did not significantly alter their values (values varied within $\pm 10\%$ range, while it is shown in **Fig. 7.12** that variations by up to 50% would not increase the fitting error by more than 3%).

There is a clear correspondence between the high-contrast part of Exp. 2 and σ_θ, the high-contrast part of Exp. 3 and σ_λ, and the region for $5 - 25°$ masks orientations in Exp. 4 and Σ_θ **(Figs. 7.10, 7.13)**, but we expected a clearer correspondence between Exp. 5 and

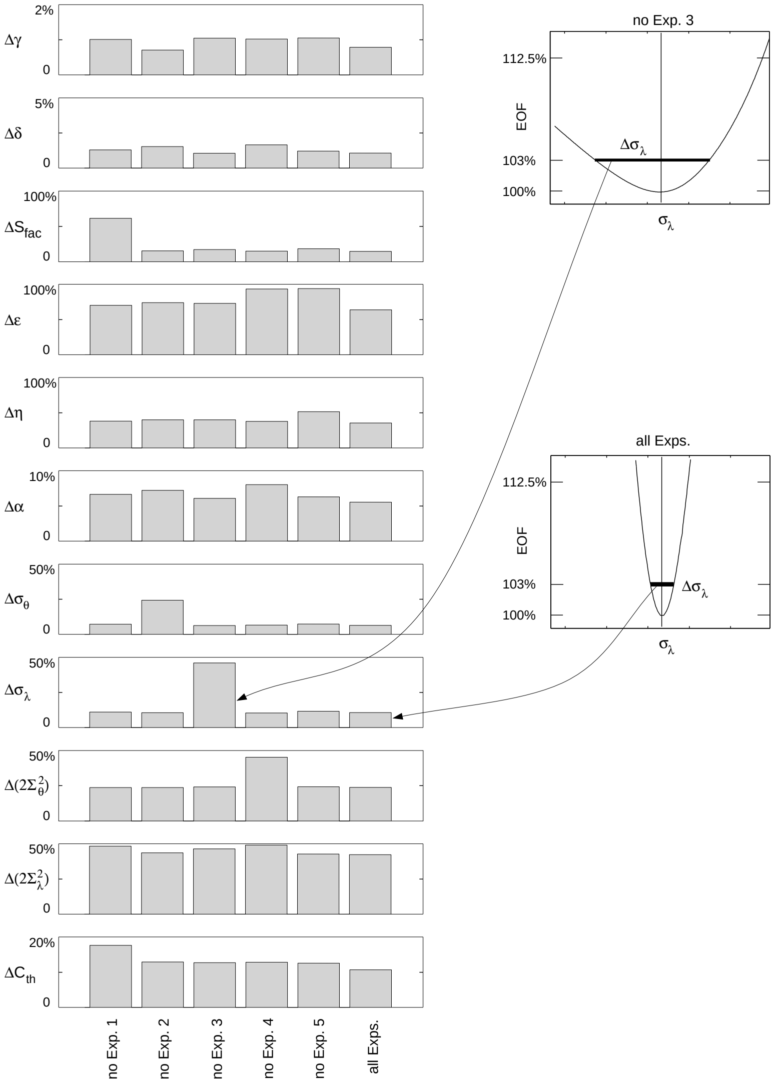


Figure 7.13: Robustness of the best-fit model parameters with partial datasets, also for the average of the datasets of observers LB, LZ and SC, like in **Fig. 7.12**. For each binary combination of experiments, parameters were varied around their best-fit values by up to ±200% in steps of 0.25%. Bars indicate the magnitude of parameter variations around their minimum which yielded an increase of 3% in the fitting error (see insets and **Fig. 7.12**). Increased bar height for a given parameter when one experiment is ignored (first five columns), relative to the height when all experiments constrain the model (last column), indicates that this experiment was responsible for constraining that parameter. For instance, the spatial period tuning width σ_λ (eighth row) became largely unconstrained when the spatial period discrimination experiment (Exp. 3) was ignored (third column and insets at right). Exp. 3 was consequently mainly responsible for constraining σ_λ.

Σ_λ. As expected from the sections in **Fig. 7.12**, the exponents γ and δ, which determine the high-contrast regime of the model, are critically constrained by all experiments. It is particularly remarkable in this case that all experiments agree very closely on the values of these parameters (this was true for all 31 possible partial dataset combinations). The fact that only Exp. 1 constrains S_{fac} is easily understood because this parameter is only critical for the very-low contrast regime of the model ($C < 5\%$), which is not present in the other experiments. Contrarily, although more strongly constrained by Exp. 1, C_{th} is also constrained by all the other experiments: Its value indeed reflects the general sensitivity of the model, and hence affects all fine discriminations. Similarly, it is not surprising that α, ϵ and η are not specifically constrained by any given experiment, as they also determine the overall sensitivity of the model for any type of discrimination.

Eigenvectors of the Hessian of the Error-of-Fit Function. Finally, we were interested in knowing which parameter *combinations* are most critical for model predictions. To this end, we computed the eigenvectors of the Hessian of the error surface in parameter space and sorted them by their associated eigenvalues (Appendix 7.10)). For all three observers, the largest eigenvalue was associated with a vector almost collinear to the contrast threshold C_{th}. This reflects the fact that C_{th} determines the overall sensitivity of the model and that its value modulates all predictions (both low- and high-contrast). The next largest eigenvalue was associated with the difference between exponents $\gamma - \delta$, which affects all high-contrast predictions and determines the asymptotic slope of the contrast response function in **Fig. 7.11a**. Intermediate eigenvalues were associated with more complicated parameter combinations and were not consistent across subjects. However, in all observers the two smallest eigenvalues were associated with the pooling widths Σ_θ and Σ_λ, respectively.

Summary. Through all of the stability analysis presented here, it appears that the best fit obtained for each observer is non-accidental and robust to small changes in model parameters. Computing envelopes of the model predictions allowed us to precisely characterize the contribution of each datapoint to the overall constraint set by our data onto our model; overall, we concluded from this analysis that every datapoint was useful in constraining the model. Examining accurately-computed cross-sections of the error-of-fit function demonstrated that this function appears well behaved around the best-fit point; this was of particular importance since our method for the computation of tolerance ranges for the best-fit parameters assumes that the error-of-fit function can be locally approximated by a hyperparaboloid. Examining how partial datasets constrained our model revealed that there was no apparent conflict between any of the five experiments in our dataset; indeed, when any one experiment was ignored, the resulting best-fit parameters were close to those obtained when using all experiments. This analysis also unraveled a number of simple correspondences between experimental datapoints and model parameters. Finally, computing the eigenvectors of the Hessian of the error-of-fit function showed us to which linear combinations of parameters the model was most or least sensitive.

7.3.2 Variants of the Model

We explore in this Section three variations around our model. These variations concern the number of filters used, the noise model, and the decision strategy.

Number of Units

How does the number of model units (60 units with 12 preferred orientations and 5 preferred spatial frequencies) affect our conclusions? Given out decision model, which combines information from all units, the number of model units should be of little consequence as long as it is not too small. The reason is that the number determines how densely the Fisher information surface is sampled, but does not alter its shape (**Fig. 7.6**). Indeed, when we increased the number of units to 270 (30 orientations and 9 spatial frequencies), we observed no significant threshold changes except for an approximately 10% reduction in masking thresholds (Exps. 4 and 5). A further increase to 1296 units (72 orientations and 18 spatial periods) produced no significant change in the results (i.e., no prediction changes by more than 5%).

Noise Model

We depart from previous models by assuming that response variance increases with the response mean ("Poisson noise," $V^2_{\lambda,\theta} = R^{\alpha}_{\lambda,\theta}$ with $\alpha \approx 1$). A simpler alternative is to assume that response variance is constant ("constant noise," $V^2_{\lambda,\theta} = \beta$) (Wilson & Gelb, 1984; Foley, 1994). Constant noise is a reasonable assumption as long as filters are independent and the contrast response follows a simple power law. In this case, the dependence of the signal-to-noise ratio $R_{\lambda,\theta}/V_{\lambda,\theta}$ on stimulus contrast, orientation, and spatial frequency turns out to be the same for constant noise and Poisson noise (Appendix 7.9). In the present case, however, filters interact through divisive inhibition and the contrast response follows a sigmoidal law. Thus, in principle, Poisson noise and constant noise are no longer interchangeable. However, when we fit the data of observer SC with a variant of the model using constant noise, we obtain a good fit as well **(Fig. 7.14)**. Nevertheless, the relatively broad envelopes around the best fit indicate that the constant noise model is less constrained by data and, in particular, that background activity (η) remains entirely unconstrained **(Table 7.2)**.

Decision Strategy

Another difference to earlier models is the decision stage. We use a maximum likelihood approach (Fisher information) whereas most previous models have used an approximation known as the Minkowski norm (Bowne, 1990). This approach assumes that threshold is

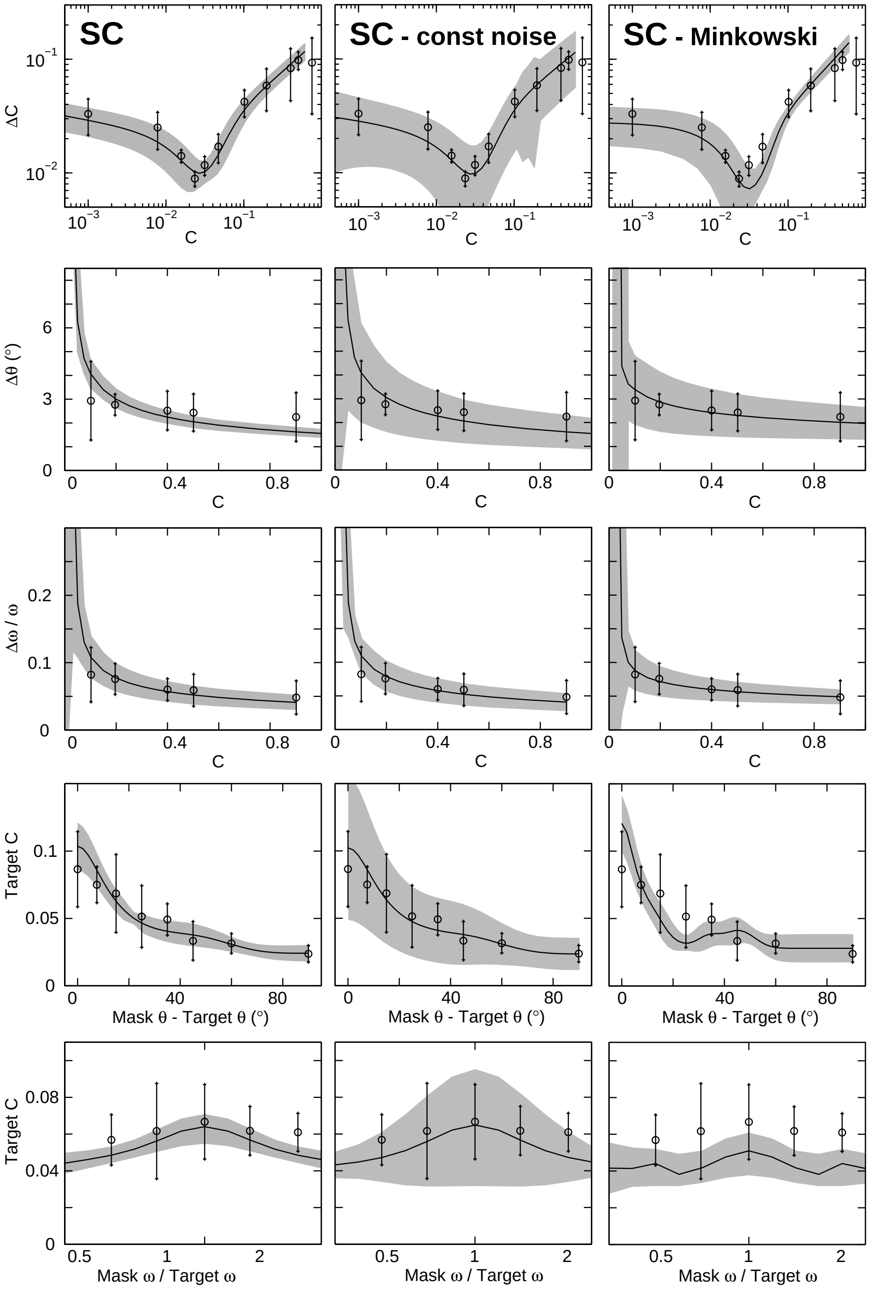


Figure 7.14: Measured and predicted thresholds for two variants of the model (observer SC). The first column shows the fit of the standard model, for comparison. The second column shows the fit of a model variant using "flat noise" instead of "proportional noise." The quality of the fit is comparable to the standard model, but is less robust for contrast masking experiments (as indicated by relatively broad grey regions). The model in the third column uses sub-optimal decision based on Minkowski norm. The fit is inferior to the standard model, in particular with respect to the contrast masking experiments. Thus, the statistically efficient decision contributes significantly to the success of the standard model.

Name	Symbol	SC	SC - const. noise	SC - Minkowski
Excitatory exponent	γ	$3.6 \pm 0.1\%$	$3.5 \pm 0.1\%$	$4.9 \pm 2\%$
Inhibitory exponent	δ	$3.0 \pm 1\%$	$3.0 \pm 1\%$	$4.0 \pm 2\%$
Noise exponent	α	$1.0 \pm 1\%$	0	$0.8 \pm 1\%$
Background activity	η	$12.3 \pm 40\%$	$15.6 \pm 4166.7\%$	$1.7 \pm 67\%$
Spatial period tuning width	σ_λ (oct)	$0.64 \pm 9\%$	$0.62 \pm 7\%$	$0.43 \pm 5\%$
Orientation tuning width	σ_θ (°)	$16.5 \pm 6\%$	$17.1 \pm 6\%$	$12.5 \pm 2\%$
Spatial period pooling width	Σ_λ (oct)	$1.2 \pm 291\%$	$0.6 \pm 138\%$	$1.4 \pm 719\%$
Orientation pooling width	Σ_θ (°)	$11.0 \pm 12\%$	$10.8 \pm 6\%$	$9.3 \pm 10\%$
Contrast detection threshold	C_{th}	$0.026 \pm 13\%$	$0.025 \pm 12\%$	$0.023 \pm 11\%$
Transducer inflexion point	S_{fac}	$0.80 \pm 8\%$	$0.80 \pm 4\%$	$1.37 \pm 6\%$
Residual fit error		17%	17%	20%
Linear spatial period FWHM*	(oct)	1.51	1.46	1.01
Linear orientation FWHM	(°)	39.1	40.3	29.4
Pooled spatial period FWHM	(oct)	0.71	0.86	0.43
Pooled orientation FWHM	(°)	26.1	28.0	18.5

* Full-width at half-maximum (FWHM) is computed as $2\sigma\sqrt{2\log(2)}$ for the linear filters, and is measured at the output of the second stage (for a grating of contrast 0.1) for the pooled filters.

Table 7.2: Best-fit model parameters for the variants of the model studied.

reached when the Minkowski norm of the differences in the mean responses to two alternative stimuli (expressed as multiples of the square root of the response variance) reaches unity (**Section 7.2.5**). To assess the importance of the decision stage for the quality of the model predictions, we fit the data of observer SC with another variant of the model using the Minkowski norm as a decision stage. **Fig. 7.14** and **Table 7.2** show the results with an exponent $Q = 3$. Although the predicted thresholds agreed well with the results of Exps. 1, 2 and 3, the predictions for Exps. 4 and 5 were rather poor.

The inferior predictive power of the Minkowski decision relates to its differential treatment of contrast masking thresholds, on the one hand, and orientation and spatial frequency discrimination thresholds, on the other hand. As illustrated in **Fig. 7.6**, contrast information is concentrated in *one*, but orientation and spatial frequency information is in *two* filter subpopulations. The Minkowski decision differs from the Fisher information decision in how information from the two subpopulations is combined: While the Minkowski discriminability increases by a factor of $2^{1/Q}$, the Fisher discriminability increases by a factor of $2^{1/2}$ (Appendix 7.9). As a result, the Minkowski decision with $Q > 2$ either *overestimates* orientation and spatial frequency thresholds or *underestimates* contrast and masking thresholds, depending on how the overall sensitivity of the model is set. Thus, the Minkowski decision contributes to the problems encountered by previous models in accounting simultaneously for different types of thresholds (Bowne, 1990).

7.3.3 Contrast Dependence of Thresholds

In an influential paper, Bowne has pointed out that a large class of models cannot explain the differential contrast dependence that the thresholds of human observers typically exhibit in the high-contrast regime (Bowne, 1990). For most observers, relative thresholds for contrast improve substantially with stimulus contrast ($\Delta C/C \propto C^{-0.3}$) while thresholds for orientation, spatial frequency, and other attributes improve little or not at all (e.g., $\Delta\theta \propto C^{-0.1}$). Contrary to this observation, many filter-based models predict that the contrast dependence of all thresholds should be the same (see also Appendix 7.9).

As our model accurately predicts the contrast dependence of all investigated thresholds, we wished to understand where Bowne's seemingly general argument fails. To this end, we manually adjusted model parameters to obtain (i) Guilford's law for contrast thresholds ($\Delta C \propto C^{0.75}$) and (ii) no contrast dependence for orientation thresholds ($\Delta\theta \propto$ *const*) **(Fig. 7.15.a,b)**. The key for obtaining this differential contrast dependence turns out to be the sigmoidal shape of the contrast response function. This sigmoidal shape distorts the orientation tuning curves in a contrast-dependent manner **(Fig. 7.15cd)**. In the region that determines orientation thresholds (*i.e.*, ±15°), the slope of the tuning function increases less with contrast than the height (as illustrated in **Fig. 7.15c**), and in the context of our decision model this shortfall suffices to keep orientation thresholds constant.

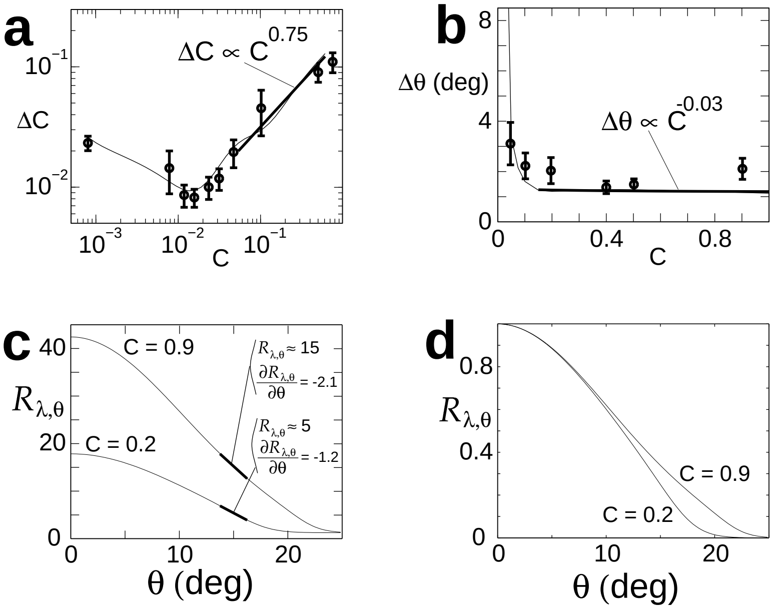


Figure 7.15: Differential contrast-dependence of thresholds. The model has been manually tuned such as to simultaneously predict increment contrast thresholds (**a**; Exp. 1) following Guilford's law ($\Delta C \propto C^{0.75}$), but contrast-independent orientation thresholds (**b**; Exp. 2) ($\Delta\theta \propto C^{-0.03}$). Looking at the internals of the model reveals that, although the unit responses, $R_{\lambda,\theta}$, increase for units of all orientations **(c)**, the increase is more pronounced at the tails of the orientation tuning curve **(d)**. As a result, the slope of the tuning curve at $\pm 15°$ increases more slowly than its height. Specifically, as contrast increases from $C = 0.2$ to $C = 0.9$ curves, $R_{\lambda,\theta}$ increases by a factor 3, whereas $\partial R_{\lambda,\theta}/\partial\theta$ increases only by a factor 1.75 (for $\theta \pm 15°$). As a result, the Fisher information with respect to contrast, which is approximately proportional to $(\partial R_{\lambda,\theta}/\partial\theta)^2/R_{\lambda,\theta}$, does not increase with C.

To see this point, consider the dependence of the Fisher information on the slope and height of the tuning function:

$$J_{\lambda,\theta} \approx \frac{(\partial R_{\lambda,\theta}/\partial\theta)^2}{R_{\lambda,\theta}} \tag{7.10}$$

For the most informative units ($\theta = \pm 15°$), the height $R_{\lambda,\theta}$ increases approximately 3-fold between contrasts 0.2 and 0.9, while the slope $(\partial R_{\lambda,\theta}/\partial\theta)$ increases only approximately 1.75-fold. As a result of the lower increase of the slope, $R_{\lambda,\theta}$ and $(\partial R_{\lambda,\theta}/\partial\theta)^2$ increase by the same factor, and the Fisher information remains the same. This explains why orientation (and, similarly, spatial frequency) thresholds exhibit so little contrast dependence in the high-contrast regime.

7.4 Discussion

7.4.1 Common Basis of Spatial Vision

To ascertain whether different aspects of spatial vision reflect the same level of visual processing, we have measured two types of thresholds with Gabor patterns presented at 4° of eccentricity. To our knowledge, this is the most extensive data set yet collected with consistent stimulus geometry and psychophysical procedure. Previous studies have tended to focus on one type of threshold at a time — contrast increment thresholds (Legge & Foley, 1980; Wilson, 1980b; Foley, 1994), orientation and spatial frequency discrimination thresholds (Hirsch & Hylton, 1982; Burbeck & Regan, 1983; Wilson & Gelb, 1984; Vogels & Orban, 1990; Bradley *et al.*, 1985; Garcia-Perez & Sierra-Vazquez, 1996), or contrast masking thresholds (Wilson *et al.*, 1983; Phillips & Wilson, 1984; Phillips & Wilson, 1984; Wilson *et al.*, 1983; Foley, 1994) — and thus were unable to address this larger question. The impetus for re-considering the basis of spatial vision at this time is provided by recent single-unit work in cat and monkey, which suggests that the behavioral thresholds in question might reflect neuronal response properties as early as primary visual cortex (Geisler & Albrecht, 1997; Carandini *et al.*, 1997; Sompolinsky & Shapley, 1997).

To test the hypothesis that spatial vision reflects a common neural basis, we employ a "consensus model" that combines components from several models of psychophysical (Wilson & Humanski, 1993; Foley, 1994; Zenger & Sagi, 1996; Thomas & Olzak, 1997) and neuronal sensitivity (Seung & Sompolinsky, 1993; Carandini *et al.*, 1997; Deneve *et al.*, 1999) to spatial patterns. To facilitate comparison with the presumed neural substrate, we substituted, whenever possible, the more "generic" components of neural network models for the more specialized components customary in psychophysical models (e.g., filter type, noise model, decision stage). However, the basic architecture of our model (e.g., filter population, divisive inhibition) is fully consistent with previous psychophysical models.

We fit the ten parameters of our model to 34 threshold measurements from five separate experiments. The fitting procedure affords an effectively exhaustive search through the ten-dimensional parameter space, and includes several randomly chosen starting points as well as two independent methods of iteration. The overall quality of fit is high, and the residual error is almost always smaller than the precision of measurement. To test the consistency of the result, we fit the model to separate data sets from three different observers. In spite of substantial differences in the threshold data, the best-fitting parameter values are almost always consistent and differ by 10% to 30% of their value between observers. To assess robustness of fit, we compute a "tolerance region" around the optimal value of each parameter, within which the overall quality of the fit degrades by less than 5% (APPENDIX 7.10). Almost all parameters were tightly constrained by data, some to within a few percent of their optimal value. The robustness of the fit is illustrated graphically by the "gray regions" in **Figs. 7.10** and **7.14**, which are generated by allowing all parameters to vary within their respective tolerance regions. In sum, our results demonstrate that a single model accounts quantitatively for all investigated thresholds, and these results are fully consistent with a common basis of spatial vision.

This conclusion is further strengthened by the high degree of interdependence between different types of thresholds. When we investigated which thresholds constrain different parts of the model, we found that most parts are constrained by all three types of thresholds, albeit not to the same degree. For example, the increase of visual responses with contrast ("transducer function") is constrained not only by experiments that vary contrast explicitly (Exps. 1, 2, and 3), but also by those that do not (Exps. 4 and 5). Similarly, the dependence of visual responses on orientation and spatial frequency ("tuning functions") is constrained both by experiments that explicitly vary orientation and spatial frequency (Exps. 2, 3, 4, and 5) and by experiments that do not (Exp. 1). Thus, the fact that our model accounts for a broad range of thresholds is a genuine result, and not the artificial consequence of adjusting a different model component to satisfy each type of constraint.

7.4.2 Relationship to Previous Models

Following previous psychophysical models (Wilson *et al.*, 1990; Thomas & Olzak, 1997), we postulate visual filters tuned to a range of orientations and spatial frequencies, but all centered at the same spatial location. However, we depart from previous models by assuming similar response characteristics for all filters (i.e., same contrast gain, background noise, width of orientation tuning, and width of spatial frequency tuning in octaves). This reduces the number of free parameters without compromising the quality of the fit for our particular data set, which is restricted to thresholds measured at a spatial frequency of 2.8cpd. Naturally, this limits the applicability of our model to spatial frequencies other than 2.8cpd. We note, however, that even our restricted data requires the inclusion of visual filters operating at multiple scales, in that a model with filters of a single scale produces substantially inferior fits, particularly for Exps. 4 and 5 (results not shown).

We chose visual filters defined by their polar response in the Fourier domain, which exhibit Gaussian tuning in orientation and spatial frequency. As functions of visual space, these filters closely resemble the multi-lobed functions employed by other models, except that they are not separable along x- and y- dimensions. We saw no particular need to retain separability, as this property is unlikely to have any functional bearing. The advantage of filters with Gaussian tuning is that they render the effect of non-linear interactions more transparent and easier to characterize. As discussed below, the *effective* tuning function is shaped by non-linear interactions between visual filters and departs from the Gaussian ideal.

Like with many recent models (Wilson *et al.*, 1990; Wilson & Humanski, 1993; Foley, 1994; Zenger & Sagi, 1996; Thomas & Olzak, 1997; Foley & Chen, 1997; Wilkinson *et al.*, 1997), we normalize the responses of visual filters by "divisive inhibition" (Albrecht & Geisler, 1991; Heeger, 1992). However, the parameters of divisive inhibition in our model depart from precedent in two notable respects: First, responses are taken to a relatively high exponent before being subjected to divisive inhibition (i.e., approximately 3.5 in the numerator and 3.0 in the denominator of the expression for divisive inhibition), as compared to the standard values of $\gamma = \delta = 2$ used in physiological models. As the exponents are among the most tightly constrained parameters of the model, this implies that before divisive inhibition takes effect visual responses are a sharply accelerating function of contrast. Second, only filters tuned to a relatively narrow range of orientations contribute to divisive inhibitions. This "near-orientation inhibition" is consistent with the findings of a number of previous studies (Phillips & Wilson, 1984; Crook *et al.*, 1997), but conflicts with the clear evidence for "cross-orientation inhibition" reported by other studies (Olzak & Thomas, 1991; Foley, 1994). The difference may be that the latter group of studies used spatially extensive stimuli, raising the possibility that "cross-orientation inhibition" may originate at more distant stimulus locations than "near-orientation inhibition." Nevertheless, the orientation pooling width Σ_θ is well constrained by our data **(Fig. 7.16)**.

The one parameter only loosely constrained by our data was the range of spatial frequencies contributing to divisive inhibition. Previous studies have advocated both a broad range (DeAngelis *et al.*, 1992; Bowen & Wilson, 1994) and a narrow range of spatial frequencies (Ramoa *et al.*, 1986; Heeger, 1992). To decide the issue, it would be necessary to collect additional data with target stimuli of other spatial frequencies.

The main difference to other models, however, is that divisive inhibition simultaneously governs all functionally relevant properties of our model so that all types of threshold measurements reflect on this part of the model. For example, divisive inhibition deter-

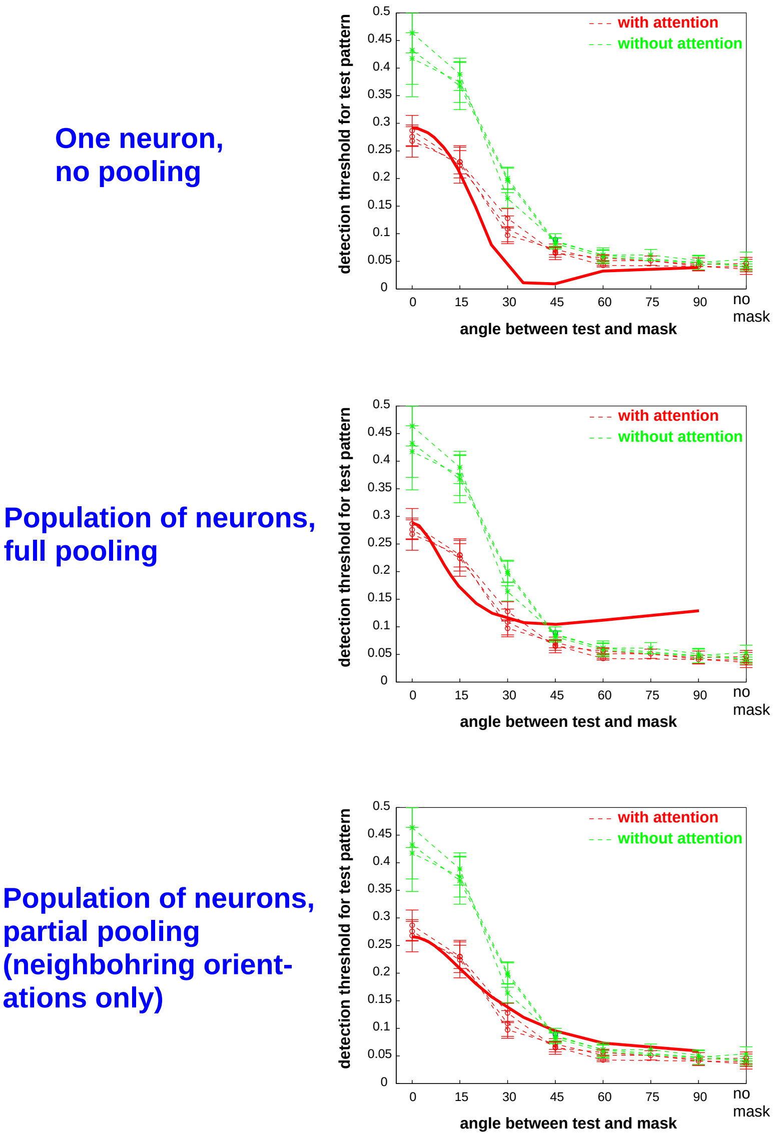


Figure 7.16: Influence of pooling across orientations on model predictions, here shown using a fit of our model to the "fully attended" (red) data of Lee *et al.* (1999) for Exp. 4. **(Top)** In the absence of any pooling across orientations, our model predicts a "dipper" when target and mask orientations differ by approximately 40°, but this feature is absent from our data. **(Middle)** With "full pooling," that is $\Sigma_\theta \to \infty$ and all filters contributing equally to the inhibitory pool, our model predicts a significant elevation of threshold even when the mask is oriented 90° away from the target. Such "cross-orientation inhibition" however is absent from the data. **(Bottom)** A good fit is obtained when only neighboring orientations participate to the inhibitory pool of each unit, as was automatically determined by our fitting procedure.

mines the sigmoidal shape of the effective contrast response function ("transducer"), which closely resembles the *ad hoc* transducer functions used by previous psychophysical models (Legge & Foley, 1980; Wilson *et al.*, 1983; Wilson, 1993; Foley, 1994; Thomas & Olzak, 1997). Similarly, divisive inhibition shapes the effective tuning for orientation and spatial frequency, which is 30% to 35% sharper than the Gaussian tuning of individual filters. Lastly, divisive inhibition determines the extent to which tuning for orientation and spatial frequency changes with contrast. Specifically, the height of the tuning curves increases more rapidly with contrast than their width decreases, ensuring that orientation and spatial frequency discrimination thresholds are relatively independent of contrast. The interconnectedness of different functional properties stands in sharp contrast to previous models, which postulate separate components to account for transducer function, tuning function, contrast-dependence of the tuning, and so on.

To predict psychophysical thresholds, it is necessary to make some assumptions about noise. We postulated that the variance of the noise distribution increases with the mean ("proportional noise" or "Poisson$^\alpha$ noise"), consistent with other recent models (Seung & Sompolinsky, 1993; Deneve *et al.*, 1999). Previous psychophysical models have employed noise of constant variance instead ("flat noise") (Legge & Foley, 1980; Wilson *et al.*, 1990; Foley, 1994; Thomas & Olzak, 1997). Although the different noise assumptions are not interchangeable in general (APPENDIX 7.9), the practical difference in the present context is small, in that model variants with proportional and flat noise afford essentially the same quality of fit. We nevertheless prefer proportional noise, because it ensures that model responses are directly comparable to neuronal responses (see below).

The final part of our model is the decision stage. Building on previous work on maximum-likelihood decisions (Seung & Sompolinsky, 1993; Pouget *et al.*, 1998; Deneve *et al.*, 1999), we have extended the Fisher-information framework to arbitrary psychophysical discriminations and sparse filter populations. This was made possible by relying on exact numerical computations rather than analytical approximations, which can be quite misleading (APPENDICES 7.8 and 7.9). The advantage of a statistically efficient decision is that it requires no task-specific assumption and thus is able to predict the relative levels of different types of thresholds. In fact, our statistically efficient decision agrees significantly better with observer thresholds than the sub-optimal decision strategies widely used in other psychophysical models ("Minkowski norm") (Thomas & Olzak, 1997; Bowne, 1990). Specifically, our optimal decision strategy yields orientation (and spatial frequency) discrimination thresholds 60% lower than the Minkowski norm, when contrast discrimination thresholds are the same, and thus corresponds much better to the threshold relationship that is actually observed. Our results bear out the theoretical analysis of Pouget and colleagues (Pouget *et al.*, 1998), who found that sub-optimal strategies yield larger discrimination thresholds. Indeed, these larger discrimination thresholds are incompatible with the observed masking thresholds.

7.4.3 Relationship to Physiology

The response of neurons in primary visual cortex typically saturates at a given contrast and is best described by a "hyperbolic ratio" function (Albrecht & Hamilton, 1982; Heeger, 1992; Geisler & Albrecht, 1995). This is at variance with our model, where responses continue to increase with contrast (as a power function with an exponent of approximately 0.5). A likely reason for the discrepancy is that model responses reflect the average response of a diverse population of neurons which saturate at different contrasts (Albrecht & Hamilton, 1982; Foley, 1994).

The tuning width of neurons in primary visual cortex of macaque has been estimated to be $20 \pm 9^\circ$ for orientation and $0.76 \pm 0.30 oct$ for spatial frequency (half-width at half-maximum, or HWHM) (DeValois *et al.*, 1982; Hirsch & Hylton, 1982; Skottun *et al.*, 1987; Geisler & Albrecht, 1997). This agrees reasonably well with the effective tuning half widths of our model units, which are 14° for orientation and 0.56oct for spatial frequency, if one considers that psychophysical performance is likely to reflect the best tuned neurons of a diverse population (Bradley *et al.*, 1985; Vogels & Orban, 1990; Britten *et al.*, 1992; Zohary *et al.*, 1994; Shadlen *et al.*, 1996). We note also that tuning widths in primary visual cortex appear to be independent of contrast (within the accuracy of physiological measurements) (Skottun *et al.*, 1987), which is once again consistent with our model (to the precision of experimental data).

The variance of neuronal responses is roughly proportional to the mean, the exact relation being a power function with an exponent of 1.1 to 1.2 (Softky & Koch, 1993; Teich *et al.*, 1996; Geisler & Albrecht, 1997). Our model agrees closely with these values, as the best-fitting exponent ranged between 1.0 and 1.3. Note, however, that our model assumes that responses vary independently from each other. This is not quite true in cortex, although the conditional covariance between visual cortical neurons is small (i.e., the correlation not due to the stimulus) (Zohary *et al.*, 1994; Gawne *et al.*, 1996). In our model, a population of units always contains more information than the most informative individual unit, because Fisher information is additive for independent noise. In cortex, however, individual neurons may encode as much information about a stimulus as the animal as a whole (Bradley *et al.*, 1985; Vogels & Orban, 1990; Britten *et al.*, 1992; Zohary *et al.*, 1994; Shadlen *et al.*, 1996). The reason for this discrepancy may lie in conditional covariance between neuronal responses, as the information of individual units would no longer be additive.

A particularly interesting issue is the neural basis of divisive inhibition. Although divisive inhibition was originally considered the result of shunting inhibition at the level of individual neurons (Carandini & Heeger, 1994), it is now simply thought to provide a convenient description of the collective behavior of neural circuits in primary visual cortex

(Somers *et al.*, 1995; Douglas *et al.*, 1995b; Carandini *et al.*, 1997; Holt & Koch, 1997; Ahmed *et al.*, 1997). The function of these circuits remains controversial and may have to do with contrast adaptation, with the sharpening of neuronal tuning, or simply with gain control (Ferster & Koch, 1987; Bolz *et al.*, 1989; Heeger, 1992; Somers *et al.*, 1995). The circuits we describe in terms of divisive inhibition are likely to be found among short-range intrinsic connections in primary visual cortex (Toyama *et al.*, 1981; Blasdel & Salama, 1986; Katz *et al.*, 1989; Crook *et al.*, 1997; Das & Gilbert, 1999). This is suggested by the functional organization of primary visual cortex and the fact that the circuits in question involve neurons with overlapping receptive fields and similar tuning properties. Recurrent excitation and inhibition within cortical columns are likely to play an important role as well (Douglas & Martin, 1991; Hata *et al.*, 1991; Douglas *et al.*, 1995b). Indeed, detailed computational models combining short-range inhibition and recurrent excitation exhibit functionalities which are very similar to divisive inhibition (*e.g.*, sharpening of orientation tuning and reducing its dependence on contrast) (Somers *et al.*, 1995; Carandini & Ringach, 1997).

Another feature of the interactions implemented in the model was that we found the width of the inhibitory pool to be narrow in the orientation domain. This property results from masking data, which does not exhibit any cross-orientation inhibition. As regards the excitatory pool, our model used a single unit. This connectivity is directly supported by studies of intrinsic connections in cat hypercolumns (e.g., (Hata *et al.*, 1991)), where excitatory connections are predominantly found between cells with same tuning, and inhibitory connections between cells with similar tuning but not between cells with orthogonal tuning (see also refs. (Ferster, 1986; Ferster, 1988)). For the time being, our model does not make a clear prediction towards narrow (Ramoa *et al.*, 1986; Heeger, 1992) or broad (DeAngelis *et al.*, 1992; Bowen & Wilson, 1994) pools in spatial period.

Although our decision stage is purely abstract, and not intended as a model of any particular level of cortical processing, it is interesting to note that neural networks are well suited to compute statistically efficient estimates of stimulus attributes such as contrast, orientation, or spatial frequency. In fact, several network architectures have been proposed as maximum-likelihood estimators (Seung & Sompolinsky, 1993; Mato & Sompolinsky, 1996). Particularly relevant to our model is a non-linear recurrent network for estimating stimulus orientation (Pouget *et al.*, 1998; Deneve *et al.*, 1999).

7.5 Outlook

In this Chapter, we have seen that our model indicates that a wide range of spatial vision thresholds reflect a single level of visual processing, most likely corresponding to primary visual cortex. Furthermore, the visual processing in question is described quite well by divisive inhibition among overlapping visual filters. Both conclusions follow from our finding that using divisive inhibition simultaneously predicts — and in return is simultaneously constrained by — contrast increment thresholds, orientation and spatial frequency discrimination thresholds, and contrast masking thresholds. The parameters of divisive inhibition, as inferred from threshold data, are tightly constrained. For the most part, they are in excellent agreement with what is known about visual processing at the level of primary visual cortex. The one exception is the range of spatial frequencies contributing to divisive inhibition, which is only weakly constrained by our data. In accounting for a wide range of behavioral thresholds, we found it important to employ a statistically efficient decision that avoids any bias in favor of one threshold or another. To this end, we described a generalized Fisher- information approach that can be adapted to arbitrary psychophysical tasks (see APPENDICES).

In the next Chapter, we will see how the present model also accounts for top-down modulation of psychophysical thresholds by visual attention (Lee *et al.*, 1997a; Itti *et al.*, 1999; Lee *et al.*, 1999b).

7.6 Appendix: Computation of Inhbibitory Constant

We derive a closed-form expression for the inhibitory constant, S, which ensures that the inflexion point of the transducer function will be located at a given stimulus contrast. In our model, this procedure is used to replace, in our results, the parameter S (which has no easy intuitive meaning if the range of activity, number, and weighting of the different units is taken into account) by a more intuitive parameter, S_{fac} (which relates the lowest contrast increment threshold in Exp. 1 to the contrast detection threshold).

The model, in its most general form where all parameters can have different values for the different scales $\lambda \in \Lambda$, is formulated as, for a stimulus of contrast c_S, period λ_S and orientation θ_S:

$$E_{\lambda,\theta} = G_\lambda c_S \exp\left(-\frac{([\theta_S - \theta]_{]-\pi/2,\pi/2]})^2}{2\sigma^2_{\theta_\lambda}} - \frac{(\log(\lambda_S/\lambda))^2}{2\sigma^2_{\lambda_\lambda}}\right) \tag{7.11}$$

and the pooled response is:

$$R_{\lambda,\theta} = \frac{(A_\lambda(E_{\lambda,\theta} + \epsilon_\lambda))^{\gamma_\lambda}}{(S_\lambda)^{\delta_\lambda} + \sum\limits_{(\lambda',\theta')\in\Lambda\times\Theta} W_{\lambda,\theta}(\lambda',\theta')\left(A_{\lambda'}(E_{\lambda',\theta'} + \epsilon_{\lambda'})\right)^{\delta_\lambda}} + \eta_\lambda \tag{7.12}$$

Note that this is the most general formulation of our model, and is slightly more complex than the simplified expression used in the body of this Chapter. For a given stimulus shape (λ_S, θ_S), we can simplify the expression for $E_{\lambda,\theta}$ as:

$$E_{\lambda,\theta} = \alpha_{\lambda,\theta} c_S \tag{7.13}$$

i.e., we absorb the tuning curves in a contrast-independent coefficient $\alpha_{\lambda,\theta}$ and only keep the linear dependence of $E_{\lambda,\theta}$ with c_S explicit. Obviously, the values of $\alpha_{\lambda,\theta}$ are different for different stimuli (λ_S, θ_S). The coefficients $\alpha_{\lambda,\theta}$ are unrelated to the noise exponent α used elsewhere in this Chapter, and should not be confused with it.

Given all the other parameters, we want to compute the set $\{S_\lambda; \lambda \in \Lambda\}$ which will yield an inflexion in the transducers for a fixed set of contrasts $\{c_{i_\lambda}; \lambda \in \Lambda\}$. To this end, we need to compute the second derivative of $R_{\lambda,\theta}$ with respect to c_S; this second derivative will be zero at the inflexion point. Note that the contrast at the inflexion point approximately corresponds to the contrast at which lowest thresholds are found in Exp. 1 (this approximate correspondence is dependent on the noise model). To compute the second derivative of $R_{\lambda,\theta}$ with respect to c_S, we write, using the shorthand notation $\forall f, f' \equiv \partial f/\partial c_S$:

$$\begin{aligned} R_{\lambda,\theta} &= \frac{N}{D} + \eta_\lambda && (7.14) \\ R'_{\lambda,\theta} &= \frac{N'}{D} - \frac{ND'}{D^2} && (7.15) \\ R''_{\lambda,\theta} &= \frac{1}{D^3}\left(D^2N'' - DND'' - 2DN'D' + 2ND'^2\right) && (7.16) \end{aligned}$$

with (caution, N and D both depend on λ and θ though this is not made explicit here in order to simplify the notations):

$$\begin{aligned} N &= (A_\lambda(\alpha_{\lambda,\theta}c_S + \epsilon_\lambda))^{\gamma_\lambda} && (7.17) \\ N' &= \gamma_\lambda A_\lambda \alpha_{\lambda,\theta} \left(A_\lambda(\alpha_{\lambda,\theta}c_S + \epsilon_\lambda)\right)^{\gamma_\lambda - 1} && (7.18) \\ N'' &= \gamma_\lambda(\gamma_\lambda - 1)A_\lambda^2\alpha_{\lambda,\theta}^2 \left(A_\lambda(\alpha_{\lambda,\theta}c_S + \epsilon_\lambda)\right)^{\gamma_\lambda - 2} && (7.19) \\ D &= (S_\lambda)^{\delta_\lambda} + \sum_{(\lambda',\theta')\in\Lambda\times\Theta} W_{\lambda,\theta}(\lambda',\theta') \left(A_{\lambda'}(\alpha_{\lambda',\theta'}c_S + \epsilon_{\lambda'})\right)^{\delta_\lambda} && (7.20) \\ D' &= \delta_\lambda \sum_{(\lambda',\theta')\in\Lambda\times\Theta} A_{\lambda'}\alpha_{\lambda',\theta'} W_{\lambda,\theta}(\lambda',\theta') \left(A_{\lambda'}(\alpha_{\lambda',\theta'}c_S + \epsilon_{\lambda'})\right)^{\delta_\lambda - 1} && (7.21) \\ D'' &= \delta_\lambda(\delta_\lambda - 1) \sum_{(\lambda',\theta')\in\Lambda\times\Theta} A_{\lambda'}^2\alpha_{\lambda',\theta'}^2 W_{\lambda,\theta}(\lambda',\theta') \left(A_{\lambda'}(\alpha_{\lambda',\theta'}c_S + \epsilon_{\lambda'})\right)^{\delta_\lambda - 2} && (7.22) \end{aligned}$$

For $D \neq 0$ (which is always true), $R''_{\lambda,\theta} = 0 \Leftrightarrow D^3R''_{\lambda,\theta} = 0$. Hence, $R''_{\lambda,\theta}$ is zero when:

$$D^2N'' - DND'' - 2DN'D' + 2ND'^2 = 0 \tag{7.23}$$

Of all the variables in the above equation, only D depends on S_λ, such that we can rewrite D as:

$$D = (S_\lambda)^{\delta_\lambda} + D_0 = U_\lambda + D_0 \tag{7.24}$$

We now solve Eq. 7.23 for U_λ, which will give us S_λ:

$$
\begin{aligned}
(U_\lambda^2 + D_0^2 + 2U_\lambda D_0)N'' - (U_\lambda + D_0)(ND'' + 2N'D') + 2ND'^2 &= 0 && (7.25)\\
N''U_\lambda^2 + (2D_0N'' - ND'' - 2N'D')U_\lambda + & && (7.26)\\
(D_0^2N'' - D_0(ND'' + 2N'D') + 2ND'^2) &= 0 && (7.27)\\
aU_\lambda^2 + bU_\lambda + c &= 0 && (7.28)\\
\Delta &= b^2 - 4ac && (7.29)\\
U_\lambda &= \frac{-b \pm \sqrt{\Delta}}{2a} && (7.30)
\end{aligned}
$$

Hence U_λ, hence S_λ. Thus, using the desired contrasts at inflexion c_{i_λ} as the value of c_S for the different scales $\lambda \in \Lambda$, we obtain the corresponding set of inhibitory constants $\{S_\lambda; \lambda \in \Lambda\}$.

Several variations of the model exist, but all differences are in the expression of the noise, which is not used here. As a consequence, we may have $\forall \lambda \in \Lambda, A_\lambda = 1$ and $\eta_\lambda = 0$, if these parameters are absorbed in the noise formulation. The computation of S_λ presented here does not change under these assumptions.

7.7 Appendix: Fisher Information in the General Gaussian Case

It has been shown by others that, for a Poisson random variable with rate ζ, Fisher information for a function depending on ζ, $f(\zeta)$ is equal to (Seung & Sompolinsky, 1993; Pouget *et al.*, 1998):

$$J(\zeta) = \frac{f'(\zeta)^2}{f(\zeta)} \tag{7.31}$$

In our model, however, we use a more general noise model, which can take any form which can be reasonably well approximated by a Gaussian. This includes the "Poisson$^{\alpha}$" noise model described in the main body of this Chapter. The use of such more general noise model requires us to derive Fisher information corresponding to that model.

Consider a Gaussian random variable x with mean μ and variance σ^2, both of which are functions of a stimulus attribute ζ. We now derive the Fisher information of x with respect to ζ. The probability of observing x given ζ, $p(x|\zeta)$, and its derivative with respect to ζ, $p'(x|\zeta)$, are:

$$p(x|\zeta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{7.32}$$

$$p'(x|\zeta) = p(x|\zeta)\left(\frac{(x-\mu)^2}{\sigma^3}\sigma' + \frac{x-\mu}{\sigma^2}\mu' - \frac{\sigma'}{\sigma}\right) \tag{7.33}$$

If we denote the expectation of $f(x)$ with respect to to $p(x|\zeta)$ with $E[f(x)]$,

$$E[f(x)] = \int_{-\infty}^{+\infty} f(x)p(x|\zeta)dx \tag{7.34}$$

the definition of Fisher information (*e.g.*, (Cover & Thomas, 1991)) yields:

$$\begin{aligned}
J(\zeta) \quad &= E\left[\left(\frac{\partial}{\partial\zeta}\log p(X|\zeta)\right)^2\right] \\
&= E\left[\frac{p'(X|\zeta)^2}{p(X;\zeta)^2}\right] \\
&= E\left[\left(\frac{(X-\mu)^2}{\sigma^3}\sigma' + \frac{X-\mu}{\sigma^2}\mu' - \frac{\sigma'}{\sigma}\right)^2\right] \\
&= \frac{1}{\sigma^6}\left[\left(\mu^2\sigma^2\mu'^2 - 2\mu^3\sigma\mu'\sigma' + 2\mu\sigma^3\mu'\sigma' + \mu^4\sigma'^2 - 2\mu^2\sigma^2\sigma'^2 + \sigma^4\sigma'^2\right)\right. \\
&\quad + \left(-2\mu\sigma^2\mu'^2 + 6\mu^2\sigma\mu'\sigma' - 2\sigma^3\mu'\sigma' - 4\mu^3\sigma'^2 + 4\mu\sigma^2\sigma'^2\right) E[x] \\
&\quad + \left(\sigma^2\mu'^2 - 6\mu\sigma\mu'\sigma' + 6\mu^2\sigma'^2 - 2\sigma^2\sigma'^2\right) E[x^2]
\end{aligned}$$

$$\begin{aligned} &+ \left(2\sigma\mu'\sigma' - 4\mu\sigma'^2\right) E[x^3] \\ &+ \left(\sigma'^2\right) E[x^4] \Big] \\ &= \frac{\mu'^2 + 2\sigma'^2}{\sigma^2} \end{aligned} \tag{7.35}$$

where μ' and σ' are the derivatives of μ and σ with respect to ζ, respectively. In the special case of $\sigma^2 = \mu^\alpha$ (Poisson$^\alpha$ noise), the Fisher information becomes

$$J(\zeta) = \frac{\mu'^2}{\mu^2}\left[\mu^{2-\alpha} + \frac{\alpha^2}{2}\right] \tag{7.36}$$

7.8 Appendix: Ideal Observer Decision in the General Gaussian Case

We consider a stimulus parameter ζ and two visual stimuli S_1 and S_2 corresponding to parameter values ζ_1 and ζ_2. When S_i is presented ($i \in \{1,2\}$), the observer builds an internal estimate of ζ_i. Because the estimation process is noisy, we describe the estimate of ζ_i by a Gaussian random variable $\mathcal{G}(\mu_i, \sigma_i)$ of mean μ_i and standard deviation σ_i. In the main body of this Chapter, we used unbiased ($\mu_i = \zeta_i$) and statistically efficient ($\sigma_i^2 = 1/\mathcal{J}(\zeta_i)$) estimation **(Fig. 7.5)**. Without loss of generality, we assume $\mu_1 \leq \mu_2$.

Although some of the results presented here are well known in some particular cases, like, for example, when $\sigma_1 = \sigma_2$ (Green & Swets, 1966), I have not found any complete treatment of the general case just exposed. Here, I consequently propose to carry out the full derivations in the general Gaussian case, which is what is required for our model where a variety of noise models can be used.

7.8.1 Yes/No task

The Yes/No paradigm uses a single stimulus presentation for each trial, consisting of either S_1 or S_2. The observer performs a forced choice between two hypotheses: In hypothesis H_A, which the experimenter presents with probability $P(H_A)$, S_1 is present, and in hypothesis H_B, presented with probability $P(H_B) = 1 - P(H_A)$, S_2 is present. The noisy observation is described by a random variable X, such that $P(X) = P(X|H_A)P(H_A) + P(X|H_B)P(H_B)$. In the Yes/No case, $P(X|H_A)$ is distributed from $\mathcal{G}(\mu_1, \sigma_1)$ and $P(X|H_B)$ from $\mathcal{G}(\mu_2, \sigma_2)$.

The ideal observer makes a maximum a posteriori (MAP) decision: He or she wants to maximize $P(H_\Gamma|X)$ with respect to $\Gamma \in \{A, B\}$, i.e., the probability that a given stimulus was present knowing the unique observation. The map decision rule hence is:

$$P(H_A|X) \underset{H_B}{\overset{H_A}{\gtrless}} P(H_B|X) \tag{7.37}$$

This is optimal in the sense that the total number of errors is minimized on average (assuming that misses and false alarms are equally penalizing). Using Bayes theorem,

$$P(H_\Gamma|X) = \frac{P(X|H_\Gamma)P(H_\Gamma)}{P(X)}, \tag{7.38}$$

we can show that MAP is equivalent to a maximum likelihood (ML) decision for this task: The MAP observer is equivalent to basing the decision on the likelihood ratio $\mathcal{L}$:

$$\mathcal{L} = \frac{P(X|H_A)}{P(X|H_B)} \underset{H_B}{\overset{H_A}{\gtrless}} \frac{P(H_B)}{P(H_A)} \tag{7.39}$$

Two decision criteria D and D' are thus derived, as the crossover points between the two weighted distributions $P(H_A)\mathcal{G}(\mu_1, \sigma_1)$ and $P(H_B)\mathcal{G}(\mu_2, \sigma_2)$, by solving for d in:

$$\frac{P(H_A)}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{(d-\mu_1)^2}{2\sigma_1^2}\right) = \frac{P(H_B)}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{(\mu_2-d)^2}{2\sigma_2^2}\right) \tag{7.40}$$

Taking the logarithm on both sides yields a quadratic equation in d, such that:

$$\begin{cases} D = \dfrac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2 - \sigma_1\sigma_2\sqrt{(\mu_1-\mu_2)^2 + 2(\sigma_1^2-\sigma_2^2)\log(P(H_B)\sigma_1/(P(H_A)\sigma_2))}}{\sigma_1^2-\sigma_2^2} \\ D' = \dfrac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2 + \sigma_1\sigma_2\sqrt{(\mu_1-\mu_2)^2 + 2(\sigma_1^2-\sigma_2^2)\log(P(H_B)\sigma_1/(P(H_A)\sigma_2))}}{\sigma_1^2-\sigma_2^2} \end{cases} \tag{7.41}$$

When $|\sigma_1 - \sigma_2| \to 0$ and $P(H_A) = P(H_B) = 1/2$, the expression for D extends by continuity to the classical result (Green & Swets, 1966) using a unique criterion $D \to (\mu_2 - \mu_1)/2$ (while $D' \to \pm\infty$, depending on the sign of $\sigma_2 - \sigma_1$). In general, however, the MAP observer uses *both* criteria to make a decision **(Fig. 7.17.a)**. Because this seems unecological, we assume in what follows that human observers only approximate the ideal MAP decision, by solely using the criterion D that is between μ_1 and μ_2; the observer then reports H_A when $x < D$ and H_B when $x > D$. Using this unique decision criterion, the probability of error is given by the weighted integral of the two tails of the distributions around the crossover point:

$$\begin{aligned} P(\text{error}) &= P(X \geq D|H_A)P(H_A) + P(X \leq D|H_B)P(H_B) \\ &= \frac{P(H_A)}{\sigma_1\sqrt{2\pi}}\int_D^{\infty} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)dx + \frac{P(H_B)}{\sigma_2\sqrt{2\pi}}\int_{-\infty}^{D} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)dx \\ &= \frac{P(H_A)}{2}\text{erfc}\left(\frac{D-\mu_1}{\sigma_1\sqrt{2}}\right) + \frac{P(H_B)}{2}\text{erfc}\left(\frac{\mu_2-D}{\sigma_2\sqrt{2}}\right) \\ \text{Performance} &= 1 - P(\text{error}) = \frac{1}{2} + \frac{P(H_A)}{2}\text{erf}\left(\frac{D-\mu_1}{\sigma_1\sqrt{2}}\right) + \frac{P(H_B)}{2}\text{erf}\left(\frac{\mu_2-D}{\sigma_2\sqrt{2}}\right) \end{aligned}$$

where $\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)dt = 1 - \text{erfc}(x)$ is the cumulative Normal distribution.

7.8.2 Two-alternative forced-choice (2AFC) task

The two-alternative forced-choice (2AFC) paradigm uses two stimulus presentations for each trial. The goal of this double presentation is to eliminate potential problems associated with systematic observer bias in the determination of D for the Yes/No task. For temporal 2AFC, the observer makes a forced-choice decision between two possible hypotheses: In hypothesis H_A, stimulus S_1 appears first, followed by S_2 (or in spatial 2AFC, S_1 is on the left and S_2 on the right); in hypothesis H_B, S_2 appears first, followed by S_1 (or S_2 is on the left and S_1

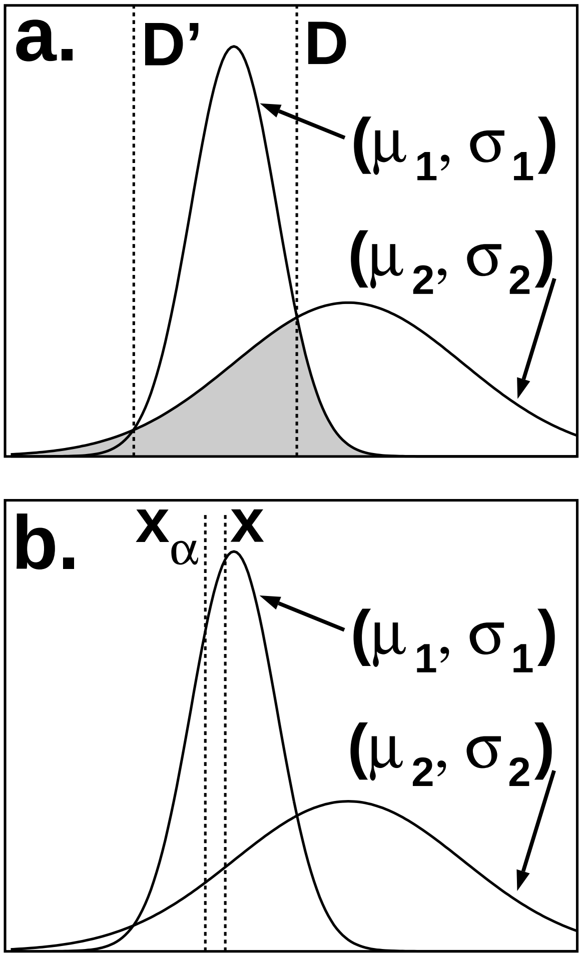


Figure 7.17: Ideal observer discrimination. **(a)** Yes/No decision: The MAP decision rule consists of deciding that the unique observation x originated from the distribution (μ_1, σ_1) iff $D' < x < D$ (and originated from (μ_2, σ_2) otherwise). Human observers most probably only approximate MAP decision by ignoring D'; their probability of error is then given by the shaded area. **(b)** 2AFC decision: Given the particular first observation x_α represented, MAP decision dictates that x_α originated from (μ_1, σ_1) and the second observation x_β originated from (μ_2, σ_2) iff $x_\beta > x_\alpha$ and $x_\beta > x$ (otherwise, x_α originated from (μ_2, σ_2) and x_β from (μ_1, σ_1)). Human observers most probably ignore the second condition, and only base their decision on the comparison between x_α and x_β (and in this example make a wrong decision, in the MAP sense, when $x_\alpha < x_\beta < x$).

on the right). The observer hence makes two observations, described by random variables X_α for the first presentation of each trial and X_β for the second presentation, such that $P(X_\alpha, X_\beta) = P(X_\alpha, X_\beta|H_A)P(H_A) + P(X_\alpha, X_\beta|H_B)P(H_B)$.

The MAP decision consists of maximizing $P(H_\Gamma|X_\alpha, X_\beta)$ with respect to $\Gamma \in \{A, B\}$. Like in the Yes/No case, MAP is equivalent to ML, such that the decision is made on the basis of the likelihood ratio $\mathcal{L}$:

$$\mathcal{L} = \frac{P(X_\alpha, X_\beta|H_A)}{P(X_\alpha, X_\beta|H_B)} \underset{H_B}{\overset{H_A}{\gtrless}} \frac{P(H_B)}{P(H_A)} \tag{7.42}$$

We assume conditional independence between X_α and X_β with respect to H_A and H_B; this simply means that the system's noise is uncorrelated between both stimulus presentations in a given trial (e.g., if the delay between both presentations is long enough for the system to reset). Hence (using the same Gaussian distributions as in the Yes/No case):

$$\begin{aligned} \mathcal{L} &= \frac{P(X_\alpha|H_A)P(X_\beta|H_A)}{P(X_\alpha|H_B)P(X_\beta|H_B)} \\ \log(\mathcal{L}) &= -\frac{(x_\alpha - \mu_1)^2}{2\sigma_1^2} - \frac{(x_\beta - \mu_2)^2}{2\sigma_2^2} + \frac{(x_\alpha - \mu_2)^2}{2\sigma_2^2} + \frac{(x_\beta - \mu_1)^2}{2\sigma_1^2} \\ &= \frac{x_\alpha - x_\beta}{2\sigma_1^2\sigma_2^2}\left(2(\sigma_2^2\mu_1 - \sigma_1^2\mu_2) + (x_\alpha + x_\beta)(\sigma_1^2 - \sigma_2^2)\right) \end{aligned}$$

The problem of solving $\log(\mathcal{L}) \underset{H_B}{\overset{H_A}{\gtrless}} \log(P(H_B)/P(H_A))$ hence yields the following rules:

$$\begin{cases} \textbf{If } x_\alpha > x_\beta, & (x_\alpha + x_\beta)(\sigma_1^2 - \sigma_2^2) \underset{H_B}{\overset{H_A}{\gtrless}} \dfrac{2\sigma_1^2\sigma_2^2 \log(P(H_B)/P(H_A))}{x_\alpha - x_\beta} - 2(\sigma_2^2\mu_1 - \sigma_1^2\mu_2) \\ \textbf{If } x_\alpha < x_\beta, & (x_\alpha + x_\beta)(\sigma_1^2 - \sigma_2^2) \underset{H_A}{\overset{H_B}{\gtrless}} \dfrac{2\sigma_1^2\sigma_2^2 \log(P(H_B)/P(H_A))}{x_\alpha - x_\beta} - 2(\sigma_2^2\mu_1 - \sigma_1^2\mu_2) \end{cases} \tag{7.43}$$

Like in the Yes/No case, it would seem unecological for human observers to use such complicated multicriteria decision rules (but note their substantial simplification when $P(H_A) = P(H_B) = 1/2$; see **Fig. 7.17.b** for an example). It is consequently assumed that humans only approximate the MAP ideal observer, by solely using the first rule, $x_\alpha \underset{H_A}{\overset{H_B}{\gtrless}} x_\beta$, irrespectively of the results from the second rule. The probability of error is then given by (using the fact that $\mathcal{G}(\mu_2, \sigma_2) - \mathcal{G}(\mu_1, \sigma_1) = \mathcal{G}(\mu_2 - \mu_1, \sqrt{\sigma_1^2 + \sigma_2^2})$):

$$\begin{aligned}
P(\text{error}) &= P(X_\alpha \geq X_\beta | H_A)P(H_A) + P(X_\alpha \leq X_\beta | H_B)P(H_B) \\
&= P(X_\alpha - X_\beta \geq 0 | H_A)P(H_A) + P(X_\alpha - X_\beta \leq 0 | H_B)P(H_B) \\
&= \frac{P(H_A)}{2}\text{erfc}\left(\frac{\mu_2 - \mu_1}{\sqrt{2(\sigma_1^2 + \sigma_2^2)}}\right) + \frac{P(H_B)}{2}\text{erfc}\left(\frac{\mu_2 - \mu_1}{\sqrt{2(\sigma_1^2 + \sigma_2^2)}}\right) \\
&= \frac{1}{2}\text{erfc}\left(\frac{\mu_2 - \mu_1}{\sqrt{2(\sigma_1^2 + \sigma_2^2)}}\right) \\
\text{Performance} &= 1 - P(\text{error}) = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\mu_2 - \mu_1}{\sqrt{2(\sigma_1^2 + \sigma_2^2)}}\right)
\end{aligned}$$

7.9 Appendix: Comparison of noise and decision models

It is instructive to compare different noise and decision models for filters with relatively simple response properties, for which the different approaches yield similar, and in some cases identical, results. Of course, this close correspondence breaks down for filters with more complex response properties, such as those used in our model. First we derive the Fisher information for Poisson and constant noise, and then compare decisions based on one filter, on all filters, and on the Minkowski norm.

7.9.1 Poisson noise

Consider a visual filter i tuned to orientation θ_i and spatial frequency ω_i whose response is a Gaussian random variable with mean R_i and variance V_i^2 given by

$$R_i = Ac^f e^{-\frac{(\theta-\theta_i)^2}{2\sigma_\theta^2}} e^{-\frac{(\omega-\omega_i)^2}{2\sigma_\omega^2}} \qquad V_i^2 = \beta R_i \tag{7.44}$$

where A is the sensitivity, f the power of the contrast dependence, σ_θ and σ_ω the tuning widths for orientation and spatial frequency, β the noise level, and c, θ, and ω are the contrast, orientation, and spatial frequency of the stimulus. Note that such a filter differs from those used in our model by neglecting background activity, by using a power function for the contrast dependence, by fixing $\alpha = 1$ and by being independent of other filters (i.e., filters with other tuning properties).

To derive the Fisher information with respect to c, θ, and ω we note that:

$$\frac{\partial R_i}{\partial c} = \frac{fR_i}{c}, \qquad \frac{\partial R_i}{\partial \theta} = -R_i\frac{\theta-\theta_i}{\sigma_\theta^2}, \qquad \frac{\partial R_i}{\partial \omega} = -R_i\frac{\omega-\omega_i}{\sigma_\omega^2} \tag{7.45}$$

Equation 7.7 yields the Fisher information with respect to attribute ζ

$$J_i^\zeta \;=\; \frac{1}{R_i^2}\left(\frac{\partial R_i}{\partial \zeta}\right)^2\left(\frac{R_i}{\beta}+\frac{1}{2}\right) \;\approx\; \frac{1}{R_i^2}\left(\frac{\partial R_i}{\partial \zeta}\right)^2\frac{R_i}{\beta}$$

$$J_i^c \approx \frac{f^2}{c^2}\frac{R_i}{\beta}, \qquad J_i^\theta \approx \frac{(\theta-\theta_i)^2}{\sigma_\theta^4}\frac{R_i}{\beta}, \qquad J_i^\omega \approx \frac{(\omega-\omega_i)^2}{\sigma_\omega^4}\frac{R_i}{\beta} \tag{7.46}$$

7.9.2 Constant noise

Identical results can be obtained for constant noise, at least for the simple visual filters considered here. Let

$$R_i = \sqrt{A} c^{f/2} e^{-\frac{(\theta-\theta_i)^2}{4\sigma_\theta^2}} e^{-\frac{(\omega-\omega_i)^2}{4\sigma_\omega^2}} \qquad V_i^2 = \frac{\beta}{4} \tag{7.47}$$

where the constants have been chosen to facilitate comparison with the case of Poisson noise. Note that the signal-to-noise ratio R_i/V_i is four times larger with constant than with Poisson noise. The derivatives for constant noise are

$$\frac{\partial R_i}{\partial c} = \frac{f R_i}{2c} \qquad \frac{\partial R_i}{\partial \theta} = -R_i \frac{\theta - \theta_i}{2\sigma_\theta^2} \qquad \frac{\partial R_i}{\partial \omega} = -R_i \frac{\omega - \omega_i}{2\sigma_\omega^2} \tag{7.48}$$

and the Fisher information with respect to attribute ζ is

$$J_i^\zeta = \frac{4}{\beta}\left(\frac{\partial R_i}{\partial \zeta}\right)^2$$

$$J_i^c = \frac{f^2}{c^2}\frac{R_i^2}{\beta} \qquad J_i^\theta = \frac{(\theta - \theta_i)^2}{\sigma_\theta^4}\frac{R_i^2}{\beta} \qquad J_i^\omega = \frac{(\omega - \omega_i)^2}{\sigma_\omega^4}\frac{R_i^2}{\beta} \tag{7.49}$$

This result is identical to the one obtained for Poisson noise.

7.9.3 Most informative filters

The threshold in a 2AFC experiment can be related to the Fisher information by letting performance equal 3/4 and reformulating Eq. 7.6 to

$$\Delta\zeta = |\zeta_1 - \zeta_2| = k\sqrt{\frac{1/J(\zeta_1) + 1/J(zeta_2)}{2}} \approx \frac{k}{\sqrt{J((\zeta_1 + \zeta_2)/2)}} \qquad k = 2erf^{-1}\frac{1}{2} \tag{7.50}$$

If we base the decision on only one filter, we can choose the most informative filters, in other words, the one with the largest Fisher information. For contrast discrimination, J_i^c is maximal when $\theta_i = \theta$ and $\omega_i = \omega$. For orientation discrimination, J_i^θ is maximal when $(\theta_i - \theta)^2 = 2\sigma_\theta^2$ and $\omega_i = \omega$ and for spatial frequency discrimination, J_i^ω is maximal when $(\omega_i - \omega)^2 = 2\sigma_\omega^2$ and $\theta_i = \theta$. The thresholds that result from a decision based on these filters can be obtained from Eqns. 7.46 or 7.49:

$$\frac{\Delta c}{c} \approx 2k\sqrt{\frac{\beta}{Af^2c^f}} \qquad \Delta\theta \approx \sqrt{2e}k\sqrt{\frac{\beta\sigma_\theta^2}{Ac^f}} \qquad \Delta\omega \approx \sqrt{2e}k\sqrt{\frac{\beta\sigma_\omega^2}{Ac^f}} \tag{7.51}$$

7.9.4 All filters

Alternatively, we may base the decision on all filters by computing the Fisher information for the entire population. For simplicity, consider a population of filters spaced $\Delta\theta$ and $\Delta\omega$ apart and covering the entire orientation/spatial-frequency plane. The total Fisher information for this population is

$$J_{tot} = \sum_{\theta_i,\omega_i} J_i = \frac{1}{\Delta\theta\Delta\omega} \sum_{\theta_i,\omega_i} J_i \Delta\theta\Delta\omega \approx \frac{\rho}{\sigma_\theta\sigma_\omega} \int\int J_i \delta\theta\delta\omega \tag{7.52}$$

where ρ is the filter density in units of $1/\sigma_\theta\sigma_\omega$. As the filter density increases, the approximation becomes more and more exact. With the help of

$$\frac{1}{\sqrt{2\pi\sigma_\theta^2}} \int \exp\left[-\frac{(\theta-\theta_i)^2}{2\sigma_\theta^2}\right] \delta\theta = 1 \qquad \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \int \frac{(\theta-\theta_i)^2}{2\sigma_\theta^2} \exp\left[-\frac{(\theta-\theta_i)^2}{2\sigma_\theta^2}\right] \delta\theta = \frac{1}{2}$$

we obtain the following expressions for the total Fisher information and the discrimination threshold:

$$J^c_{tot} \approx \frac{\rho}{\sigma_\theta\sigma_\omega} \int\int J^c_i \delta\theta\delta\omega = \frac{2\pi\sigma_\theta\sigma_\omega A f^2 c^f \rho}{\beta c^2} \qquad \frac{\Delta c}{c} \approx \sqrt{\frac{2}{\pi}} k \sqrt{\frac{\beta}{A f^2 c^f \rho}} \tag{7.53}$$

$$J^\theta_{tot} \approx \frac{\rho}{\sigma_\theta\sigma_\omega} \int\int J^\theta_i \delta\theta\delta\omega = \frac{2\pi\sigma_\omega A c^f \rho}{\beta\sigma_\theta} \qquad \Delta\theta \approx \sqrt{\frac{2}{\pi}} k \sqrt{\frac{\beta\sigma_\theta^2}{A c^f \rho}} \tag{7.54}$$

$$J^\omega_{tot} \approx \frac{\rho}{\sigma_\theta\sigma_\omega} \int\int J^\omega_i \delta\theta\delta\omega = \frac{2\pi\sigma_\theta A c^f \rho}{\beta\sigma_\omega} \qquad \Delta\omega \approx \sqrt{\frac{2}{\pi}} k \sqrt{\frac{\beta\sigma_\omega^2}{A c^f \rho}} \tag{7.55}$$

Note that the contrast threshold based on all filters is smaller than the threshold based on the most informative filter, by a factor of $1/\sqrt{2\pi}$. The difference in orientation and spatial frequency thresholds is even larger, with a factor of $1/\sqrt{e\pi}$.

7.9.5 Minkowski norm

The Minkowski norm has been one of the most popular ways to model a perceptual decision (Quick, 1974; Bowne, 1990). As thresholds reflect the magnitude of the response difference to two stimulus alternatives, one expects that thresholds will be proportional to the derivative of the response R_i with respect to a stimulus attribute ζ. Specifically, one may postulate that threshold is reached when the response difference ΔR_i equals the standard deviation V_i of the response:

$$\Delta R_i = \left| \frac{\partial R_i}{\partial \zeta} \right| \Delta\zeta = V_i \qquad \frac{\Delta R_i}{V_i} = \frac{1}{V_i} \left| \frac{\partial R_i}{\partial \zeta} \right| \Delta\zeta = 1 \tag{7.56}$$

When the decision is based on multiple filters, threshold is reached when the Minkowski norm of the ratios $\Delta R_i / V_i$ reaches unity:

$$1 \quad = \quad \left[\sum_i \left| \frac{\Delta R_i}{V_i} \right|^Q \right]^{1/Q} \tag{7.57}$$

The ratios $\Delta R_i / V_i$ stand in a simple relation to the Fisher information, which is obtained both for constant noise and for Poisson noise:

$$\frac{\Delta R_i}{V_i} \quad = \quad \frac{1}{\sqrt{\beta}} \left| \frac{\delta R_i}{\delta \zeta} \right| \Delta\zeta \quad = \quad \sqrt{J_i^\zeta} \Delta\zeta \tag{7.58}$$

In the special case of Q, the Minkowski norm predicts the same thresholds as does Fisher information, except for a proportionality constant:

$$1 \quad = \quad \left[\sum_i \left(\frac{\Delta R_i}{V_i} \right)^2 \right]^{1/2} \quad = \quad \Delta\zeta \left[\sum_i J_i^\zeta \right]^{1/2} \qquad \Delta\zeta \quad = \quad \frac{1}{\sqrt{J_{tot}^\zeta}} \tag{7.59}$$

In the general case ($Q > 2$), however, the predictions of the Minkowski norm differ from those of Fisher information. This is particularly true when information is distributed over many filters. To see this, consider a population of N filters with identical $\Delta R_i / V_i = \sqrt{J_i^\zeta} = \sqrt{J_\zeta}$:

$$\Delta\zeta_{Minkowski} = \left[\sum_i \left(\frac{\Delta R_i}{V_i} \right)^Q \right]^{-1/Q} = N^{-1/Q} \left[J_\zeta \right]^{-1/2} \tag{7.60}$$

$$\Delta\zeta_{Fisher} = \left[\sum_i J_i^\zeta \right]^{-1/2} = N^{-1/2} \left[J_\zeta \right]^{-1/2} \tag{7.61}$$

The factor by which the Minkowski norm overpredicts thresholds, $N^{1/2-1/Q}$, grows with increasing N if $Q > 2$.

7.10 Appendix: Formal Analysis of Model Fits

The robustness and stability of model fits can be analyzed with the help of two approximations: a linear approximation of the threshold function and a quadratic approximation of the fit error function. These approximations allow us to derive analytical expressions for (i) the region around the optimal value to which each parameter is constrained by data and (ii) the varying extents to which different data points constrain the model.

Let X be the vector of ten model parameters, and X_0 the best-fit value of these parameters. Each threshold prediction $t_i(X)$ for each datapoint i is linearized around X_0 as:

$$t_i(X) = t_i(X_0) + J_{t_i}^\top (X - X_0) \tag{7.62}$$

where J_{t_i} is the Jacobian of t_i at X_0. Similarly, the fit error to the data $e(X)$ is approximated by the second-order formulation:

$$e(X) = e(X_0) + J_e^\top (X - X_0) + (X - X_0)^\top H_e (X - X_0) \tag{7.63}$$

where J_e is the Jacobian and H_e the Hessian of e at X_0. At the minimum of e in X_0, we know that $J_e = 0$ and H_e is symmetric and positive.

To determine the tolerance range within which model parameters are constrained by data, we need to project the subspace $e(X) \leq e(X_0) + \epsilon$ onto each parameter's axis (see (Press *et al.*, 1992), Chap. 15 and **Fig. 7.18**). This projection yields a measure of the maximum variation obtained for each parameter when all parameters are allowed to arbitrarily vary while keeping the fit error within ϵ of the best fit error. We can write this projection problem as a constrained optimization problem: Denoting E_j the basis vector corresponding to the axis of parameter j, we want to extremize:

$$E_j^\top X \quad \text{subject to} \quad (X - X_0)^\top H_e (X - X_0) \leq \epsilon \tag{7.64}$$

To carry out this extremization, we write the Lagrange multiplier for this problem:

$$E_j^\top X - \lambda \left((X - X_0)^\top H_e (X - X_0) - \epsilon \right) \tag{7.65}$$

Differentiating this expression, successively with respect to each parameter, yields a system of equations which we can write in matrix form:

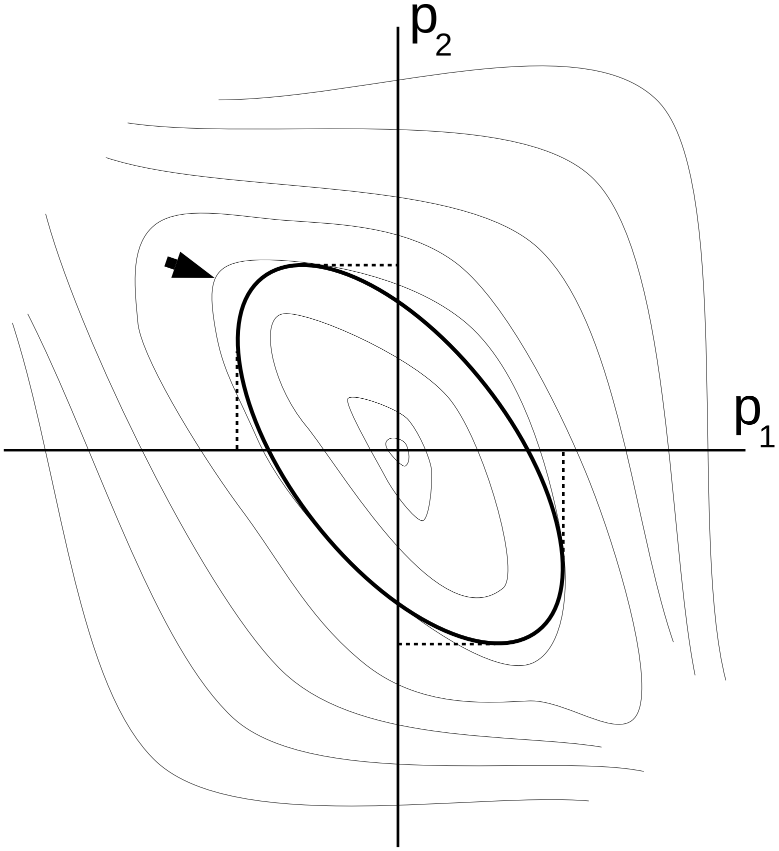


Figure 7.18: Approximation of the error surface near the point of best fit, $X_(0)$. The error surface is approximated by a paraboloid based on the Hessian matrix at X_0. With this approximation, all points at which the fit error $e(X) \leq e(X_0) + \epsilon$ are contained by an ellipsoid (thick line). The actual iso-contour at which $e(X) \leq e(X_0) + \epsilon$ is indicated by the arrow. For each parameter, the "tolerance range" within which $e(X) \leq e(X_0) + \epsilon$ is obtained by projecting the ellipsoid onto the associated axis (dashed lines).

$$E_j - 2\lambda H_e(X - X_0) = 0 \tag{7.66}$$

Using this equation and the expression for the constraint, we solve for λ and then for X:

$$X = X_0 \pm \sqrt{\frac{\epsilon}{E_j^\top H_e^{-1} E_j}} H_e^{-1} E_j, \tag{7.67}$$

hence

$$E_j^\top X = E_j^\top X_0 \pm \sqrt{\epsilon E_j^\top H_e^{-1} E_j} \tag{7.68}$$

and the tolerance range for parameter j (given in **Table 7.1**) is defined as the difference between $E_j^\top X_0$ and the maximum or minimum of $E_j^\top X$:

$$\sigma_j = \sqrt{\epsilon E_j^\top H_e^{-1} E_j} \tag{7.69}$$

To determine the relative importance of different data points in constraining the model, we need to investigate the extent of the variations in each threshold prediction $t_i(X)$ when X is allowed to arbitrarily vary while ensuring $e(X) \leq e(X_0) + \epsilon$. Since H_e is symmetric and positive, we can transform, through diagonalization of H_e, this quadratic constraint into a simple spherical inequality: We first decompose H_e as $H_e = VDV^{-1}$, where D is diagonal and V is an orthogonal change-of-basis matrix. Because $H_e \geq 0$, all its eigenvalues are positive, such that:

$$H_e = VD^{1/2}D^{1/2}V^{-1} \tag{7.70}$$

Additionally, because V is orthogonal, $V^{-1} = V^\top$. We can now rewrite the constraint as:

$$(X - X_0)^\top VD^{1/2}D^{1/2}V^\top(X - X_0) \leq \epsilon \tag{7.71}$$

$$\left(D^{1/2}V^{\top}(X - X_0)\right)^{\top}\left(D^{1/2}V^{\top}(X - X_0)\right) \leq \epsilon \tag{7.72}$$

$$W^{\top}W \leq \epsilon, \text{ i.e.,} \qquad ||W||^2 \leq \epsilon, \text{ with} \qquad W = D^{1/2}V^{\top}(X - X_0) \tag{7.73}$$

Consequently, the region of the parameter space where $e(X) < e(X_0)+\epsilon$ is the interior of the hypersphere $||W||^2 = \epsilon$. In order to compute the envelopes of all model predictions when the parameters are inside this hypersphere, we need to find the minimum and maximum of each $t_i(X)$ subject to the constraint $||W||^2 \leq \epsilon$. Since we locally approximate t_i by the linear form $t_i(X) = t_i(X_0) + J_{t_i}^{\top}(X - X_0)$, we trivially know that the extrema of this linear form will be obtained for X such that $||W||^2 = \epsilon$; indeed, we simply write:

$$t_i(X) = t_i(X_0) + J_{t_i}^{\top}V^{-\top}D^{-1/2}D^{1/2}V^{\top}(X - X_0) \tag{7.74}$$

$$t_i(X) = t_i(X_0) + K^{\top}W \tag{7.75}$$

with $K = D^{-1/2}V^{\top}J_{t_i}$, such that the extremization problem becomes that of finding the minimum and maximum of $t_i(X_0) + K^{\top}W$ subject to $||W||^2 \leq \epsilon$. Because we have reduced the function to extremize to a simple dot product between two vectors, we now see that these extrema will be obtained when W is collinear with K, and more precisely when:

$$W = \pm\sqrt{\epsilon}\frac{K}{||K||} \tag{7.76}$$

Finally, for the two values of W which extremize t_i, we can compute the minimum and the maximum of the model prediction for datapoint i from Eq. 7.75:

$$t_i^{\text{extr.}} = t_i(X_0) \pm \sqrt{\epsilon}||K|| \tag{7.77}$$

With this method, we can estimate the extrema of the range of threshold predictions obtained from arbitrary parameter values X within a certain neighborhood of the best-fit point X_0 ($e(X) \leq e(X_0) + \epsilon$). The accuracy of this estimate is limited only by the accuracy of the (very reasonable) approximations for t_i and e.

Chapter 8 Modeling Attentional Modulation of Spatial Vision

8.1 Overview

Although attention profoundly alters visual perception (Helmholtz, 1850/1962; James, 1890/1981; Pashler, 1997), it is not equally important to all aspects of vision. For example, attention is of little or no help to many detection tasks (e.g., detecting a luminance increment (Müller & Findlay, 1987b; Downing, 1988)) and the degree to which it benefits discrimination tasks varies widely with the discriminated attribute (e.g., discriminating color, orientation, form (Treisman & Gormican, 1988; Cheal & Lyon, 1994; Braun & Julesz, 1998); see the general introduction to this Part of the thesis for more introductory and background material on attentional modulation). In this Chapter, we start by summarizing experimental results (Lee, 1999) on how attention alters thresholds for discriminating contrast, orientation, and spatial frequency of simple patterns, and for detecting one pattern in the presence of another, superimposed pattern of different orientation or spatial frequency (Lee *et al.*, 1997b; Lee *et al.*, 1999a). Together, these measurements characterize the visual mechanisms that underlie basic pattern vision.

We then report that attention modulates the response normalization that seems to underlie basic pattern vision. To achieve this computational account, we use the model presented in the previous Chapter and apply it to the data of Lee *et al.* First we show that the model is capable of accurately reproducing the observations in the fully attended condition, as well as in the poorly attended condition. After these two separate fits, we manually explore a number of alterations of the model, using manipulations which have been proposed by several groups to account for a variety of attentional modulation observations. We conclude from these preliminary manipulations that only when attention alters the strength of interactions is the correct modulatory effect on thresholds reproduced by the model. We confirm this result by running joint optimizations, in which two versions of the model, which only differ by a subset of model parameters, are jointly fit to the double datasets.

8.2 Attentional Modulation of Spatial Vision Thresholds

We briefly describe the experimental method used by Lee *et al.* (1997b; 1999a) to measure attentional modulation of human psychophysical thresholds, and show their results. These experiments have been extensively described in Dr. Lee's Ph.D. thesis (Lee, 1999). We consequently refer the interested reader to Dr. Lee's thesis for further details.

8.2.1 Attentional Manipulation

Although visual thresholds are usually measured when stimuli are fully attended, here a concurrent task was used to establish thresholds when stimuli are at best poorly attended

(Braun & Sagi, 1990; Braun, 1994; Braun & Julesz, 1998). The concurrent task in question forces observers to withdraw attention from peripheral stimuli and to focus on stimuli near fixation **(Fig. 8.1)**. This psychophysical manipulation is highly effective and causes substantial perceptual deficits in the periphery, similar to the deficits obtained after a lesion in visual cortical area V4 of the monkey (Braun, 1994). However, the perception of peripheral stimuli is not entirely abolished. Practiced observers enjoy a significant residual vision outside the focus of attention and render reliable threshold judgments about peripheral stimuli, especially when the display is uncluttered and contains only a few salient stimuli (Braun & Julesz, 1998; Braun, 1998a).

Observers discriminated contrast, orientation, or spatial frequency of a luminance-modulated pattern appearing at varying locations of 4° eccentricity ("peripheral target") **(Fig. 8.1.A)**. To draw attention away from this pattern, observers were asked to discriminate whether five shapes near fixation ("central targets") were the "same" or "different." When observers carried out both tasks, they concentrated attention on the central task, which they were instructed to consider the primary task, and thus left the peripheral target "poorly attended" (double-task thresholds). In contrast, when observers viewed the same display but performed only the peripheral task, they "fully attended" to the peripheral target (single-task thresholds). The comparison of single- and double-task thresholds reveals if and how attention alters visual perception.

8.2.2 Stimulus Generation and Experimental Paradigm

The experimental setup was virtually identical to the one used for our experiments; indeed, our experiments were modeled after those of Lee *et al.* A number of differences, however, existed: A dual-task paradigm was used to manipulate attention; the tasks consisted of a Yes/No discrimination, because the 2AFC paradigm used in our experiments was impractical under dual-task conditions; consequently, the simple Gabor stimuli used in our experiments were replaced by stimuli better suited to the Yes/No discrimination.

Stimuli were generated on a Silicon Graphics Indigo with a 1280 × 1024 pixels color monitor. Viewing was binocular at $\approx$ 120 cm distance, such that $1^\circ \approx 80$ pixels. Room luminance was $3cd/m^2$, average screen luminance was $30cd/m^2$, with linear increments of $0.07cd/m^2$ obtained by Gamma correction and "color bit stealing" (Tyler, 1997). Central targets appeared at $0 - 0.8^\circ$ eccentricity and measured 0.4° across. Peripheral targets appeared at 4° eccentricity, in a circular aperture of 1.5° (**Fig. 8.1** shows timing information). They were either sinusoidal gratings **(Fig. 8.2.B,C)** or vertical stripes whose luminance profile was given by the 6th derivative of a Gaussian **(Fig. 8.2.A,D,E)**. Mask patterns were generated by superimposing 100 Gabor filters, positioned randomly within the circular aperture **(Fig. 8.2.A,D,E)**. When the spatial frequency was not varied, it was $4cpd$ (vertical stripes in **Fig. 8.2.A,D,E**; sinusoidal gratings in **Fig. 8.2.B**; superimposed

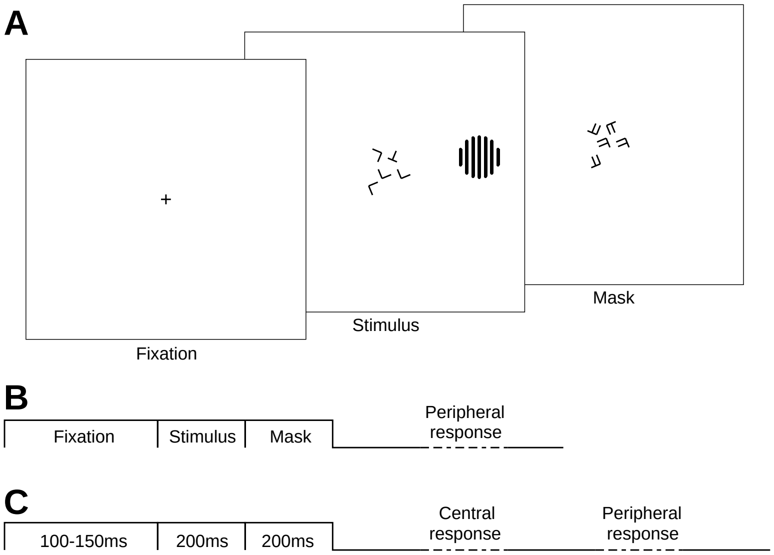


Figure 8.1: Measurement of visual thresholds with either full or poor attention. **(A)**. Sequence of fixation, stimulus, and mask displays (schematic). Observers fixate the center of all displays. The stimulus comprises a central and a peripheral component, which appear at varying locations of constant eccentricity. The central component consists of 5 Ts and/or Ls ("central targets") and observers report "same," (*i.e.,* 5 Ts or 5 Ls) or "different," (*i.e.,* 4 Ts + 1 L or 4 Ls + 1 T). The peripheral component consists of the luminance-modulated patterns shown in **Fig. 8.2.A-E** ("peripheral target"). For example, the peripheral component might be a grating pattern of vertical or tilted orientation, in which case observers would report "vertical" or "tilted." The mask display limits visual persistence of central targets. **(B)** Single-task (peripheral target 'fully attended'): observers fixate the center but respond only to the peripheral task (see **Fig. 8.2**). **(C)** Double-task (peripheral target 'poorly attended'): observers fixate the center and respond first to the central task and second to the peripheral task.

Gabors in **Fig. 8.2.A,D**). The mask contrast was 0.5 when it was constant **(Fig. 8.2.D,E)**. Thresholds were established with an adaptive staircase method (80 trials per block), *i.e.*, by adjusting target contrast, orientation, or spatial frequency in each trial according to the success or failure of previous trials (*e.g.*, (Watson & Pelli, 1983)). The figures show results from two naive observers. Symbols represent the average threshold across observers (between 12 and 20 blocks of trials per symbol). Error bars represent the average of the standard deviations computed for each observer. In the double-task situation, observers were required to match or exceed a certain level of central performance (the level achieved when the central task is carried out alone). Approximately 15% of double-task blocks were discarded because of poor central performance. In both single- and double-task situations, observers fixated the display center, ensuring identical visual stimulation. The brief presentation effectively precluded shifts of fixation towards the peripheral target (Fischer *et al.*, 1993).

8.2.3 Experimental Results

The results of Lee *et al.* can be summarized as follows. When peripheral targets are fully attended, contrast detection thresholds (for zero mask contrast) are about 20% lower, and contrast discrimination thresholds (for mask contrasts greater than zero) are about $40-50\%$ lower than when peripheral targets are poorly attended **(Fig. 8.2.A)**. In addition, the *decrease* of the discrimination threshold as mask contrast increases from zero (dipper) is evident only when targets are fully attended. Note that the target position varies from trial to trial (in order to forestall eye movements) and that positional uncertainty of this kind is known to reduce the dipper (Palmer, 1995; Solomon *et al.*, 1997; Foley & Schwarz, 1998). Therefore, it is possible that our data underestimates the depth of the dipper.

The effects of attention on spatial frequency and orientation discrimination are even more pronounced **(Fig. 8.2.B,C)**. Spatial frequency thresholds are about 60% lower and orientation thresholds about 70% lower when peripheral targets are fully attended, compared to when they are poorly attended. Note that both types of thresholds remain essentially constant for contrast values above 20%.

Interactions between superimposed stimuli of different orientation or spatial frequency (target and mask; **Fig. 8.2.D,E**) are also altered by attention. When target and mask have similar orientation or spatial frequency, attention lowers the maximal threshold by about 50% (consistent with **Fig. 8.2.A**, mask contrast 0.5). As target and mask become progressively more different, fully and poorly attended thresholds decrease towards the same baseline level. The baseline is comparable to thresholds without mask (**Fig. 8.2.A**, mask contrast 0.0), indicating minimal interactions between targets and masks of very different orientation or spatial frequency.

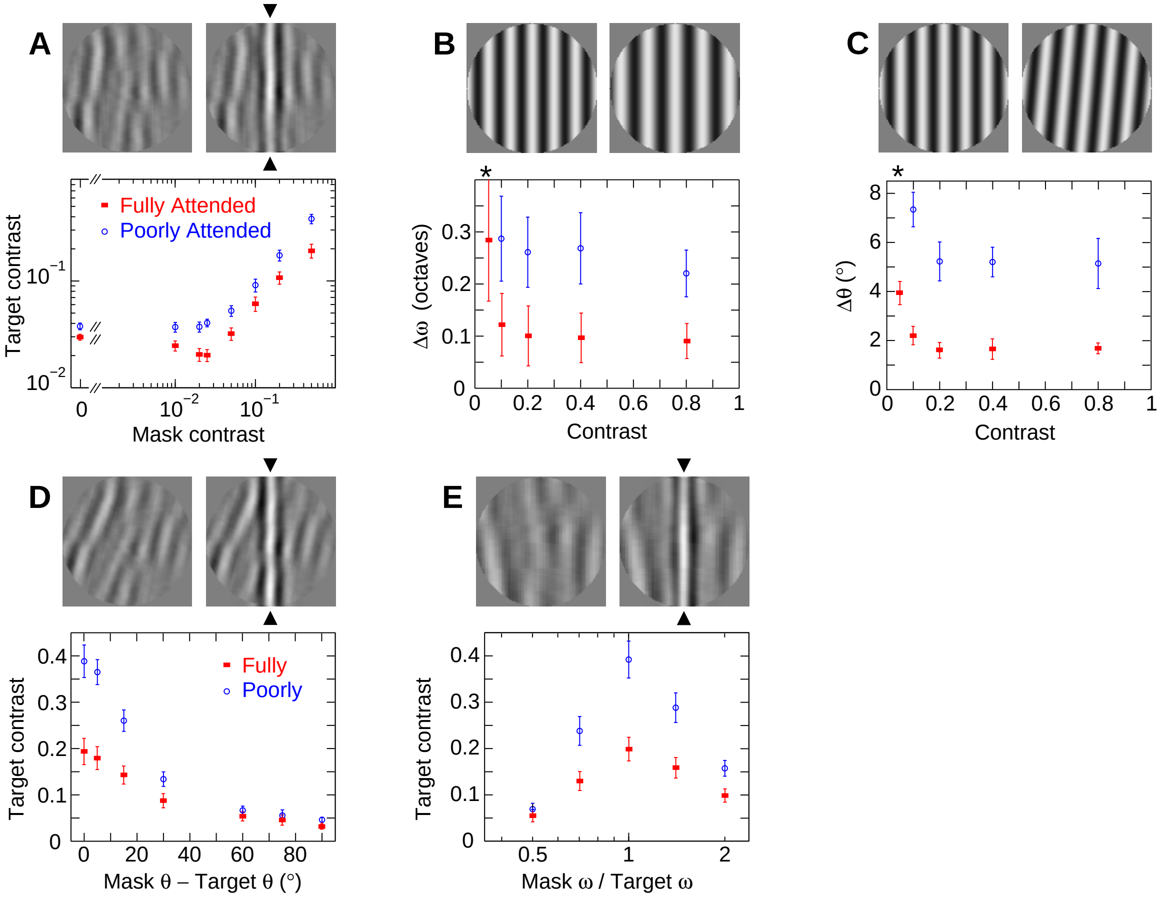


Figure 8.2: Attentional modulation of spatial vision thresholds, as observed by Lee *et al.* (1997; 1999). The experiments are identical to those presented in **Fig. 6.3** and Chapter 6, except for the following differences: First, a dual-task paradigm was used to manipulate attention, and yielded data for both a "fully attended" and a "poorly attended" conditions. Second, the tasks consisted of a Yes/No discrimination, because the 2AFC paradigm used in our experiments was impractical under dual-task conditions. Third, the simple Gabor stimuli used in our experiments were replaced by slightly more complex stimuli, better suited to the Yes/No discrimination. **(A)** Increment contrast discrimination (our Exp. 1); **(B)** spatial frequency discrimination (our Exp. 3); **(C)** orientation discrimination (our Exp. 2); **(D)** contrast masking, variable mask orientation (our Exp. 4); and **(E)** contrast masking, variable mask period (our Exp. 5).

8.2.4 Implications for Modeling

The observed attentional modulation appears complex and task-dependent. Indeed, rather small attentional effects are found on contrast discrimination, while very large effects are found on orientation and spatial frequency discrimination. Furthermore, the contrast masking experiments suggest a non-trivial selective enhancement of some thresholds by attention, while others remain unaffected, depending upon the precise configuration of target and mask stimuli.

8.3 Computational Account of Attentional Modulation

In this Section, we attack the challenging problem of trying to use our model to derive a simple, unified computational understanding of these seemingly complex experimental results. We start by deriving simple computational constraints from a simplified model, and then use the detailed model presented in the previous Chapter to provide a detailed computational account of the observed data.

8.3.1 Simplified Model

As extensively discussed in the previous Chapter, the visual thresholds measured here are thought to reflect the activity of a population of "noisy filters" selective for stimuli of different orientations and spatial frequencies (Wilson, 1980a; Legge & Foley, 1980; Watt & Morgan, 1985). In the simplified model, we define a filter tuned to orientation θ and spatial period λ by:

$$E_{\lambda,\theta}(C_S, \lambda_S, \theta_S) = C_S A \exp\left(-\frac{(\log(\lambda_S) - \log(\lambda))^2}{2\sigma_\lambda^2} - \frac{(\theta_S - \theta)^2}{2\sigma_\theta^2}\right) + B \qquad (8.1)$$

that is, in the same manner as previously presented **(Eq. 7.1)**, except for the addition of a constant background activity, B. As we saw previously, when the properties of such filters are inferred from behavioral threshold measurements, they tend to match the response properties of neurons in visual cortical areas V1 and V2 (DeValois & DeValois, 1988; Wilson *et al.*, 1990; Geisler & Albrecht, 1997). Accordingly, each visual filter is thought to correspond to a population of visual cortical neurons tuned to a particular orientation and spatial frequency.

Can the observed effects of attention be understood simply as a change in the properties of individual visual filters? To answer this question, we first examine the case in which filters are independent, so that the output of each filter, $R_{\lambda,\theta}$, is a monotonic (and perhaps non-linear) function of its linear response, $E_{\lambda,\theta}$. We also assume that the variance of the filter output, $V^2_{\lambda,\theta}$, is given by:

$$V^2_{\lambda,\theta} = \beta(R_{\lambda,\theta} + \epsilon) \tag{8.2}$$

where β is the "light noise" and ϵ is the "dark noise." This provides a good approximation of the response variance of visual cortical neurons (Geisler & Albrecht, 1997). Note that this noise model is slightly different from the Poisson$^\alpha$ model used in the previous Chapter; the main motivation for such difference is the better theoretical tractability of the present formulation, which we propose to examine here for independent filters.

Given these assumptions, the observed 20% difference between contrast detection thresholds with full and poor attention implies that either the gain A decreases or the light noise β increases by about 20%. However, the observed 60% to 70% difference in spatial frequency and orientation thresholds requires a far larger change in A and/or β (about 80%).

Indeed, if the "transducer function" $R_{\lambda,\theta} = t(E_{\lambda,\theta})$ is linear over small ranges of contrast, and if the decision between stimulus alternatives is statistically efficient, we can derive simple proportionality relationships for the contrast detection threshold, ΔC_{det}, and the orientation and spatial period discrimination thresholds, $\Delta\theta_{dis}$ and $\Delta\lambda_{dis}$:

$$\Delta C_{det} \propto \frac{\beta}{A} \tag{8.3}$$

$$\Delta\theta_{dis} \propto \sigma_\theta \sqrt{\frac{\beta}{Ac_s}} \qquad \Delta\lambda_{dis} \propto \sigma_\lambda \sqrt{\frac{\beta}{Ac_s}} \tag{8.4}$$

These relations indicate the extent to which filter parameters such as β/A, σ_θ, and σ_λ must change to produce a given change in thresholds. Note that not all thresholds depend on the same parameters. This model and the above expressions hence implies that, if we assume only a change in gain A or light noise β, we find the inconsistency mentioned: The change derived from the contrast detection thresholds is far too small to account for the observed differences in orientation or spatial frequency thresholds.

To reconcile our observations about different thresholds, already based on this simplified model, we therefore have to assume that attention alters not only the gain or noise of visual filters, but also the tuning widths, σ_θ and σ_λ, for orientation and spatial period.

8.3.2 Detailed Model

In this Section, we use the detailed model presented in the previous Chapter, with the exception that the same noise model as above is used, in order to simplify comparisons between the simplified and full models:

$$V^2_{\theta,\omega} = \beta(R_{\theta,\omega} + \epsilon) \tag{8.5}$$

As a consequence of this new noise model, we can eliminate the linear gain coefficient A from our model, which is now included in the noise coefficient β. Finally, an additional parameter is added, like in the simplified model, which expresses background activity at the linear filtering stage, B.

We start by showing that our model is capable of separately reproducing the observed data in either attentional condition. This result indicates that the model components and degrees of freedom are suited to explaining not only, as extensively studied in the previous Chapter, fully attended thresholds, but also poorly attended thresholds. This is an important result, which is not a matter of course as the internal characteristics of the best-fit models differ substantially; in particular, we find a significant difference in the noise variance, as well as in the power-law and divisive inhibition exponents, between both attentional conditions.

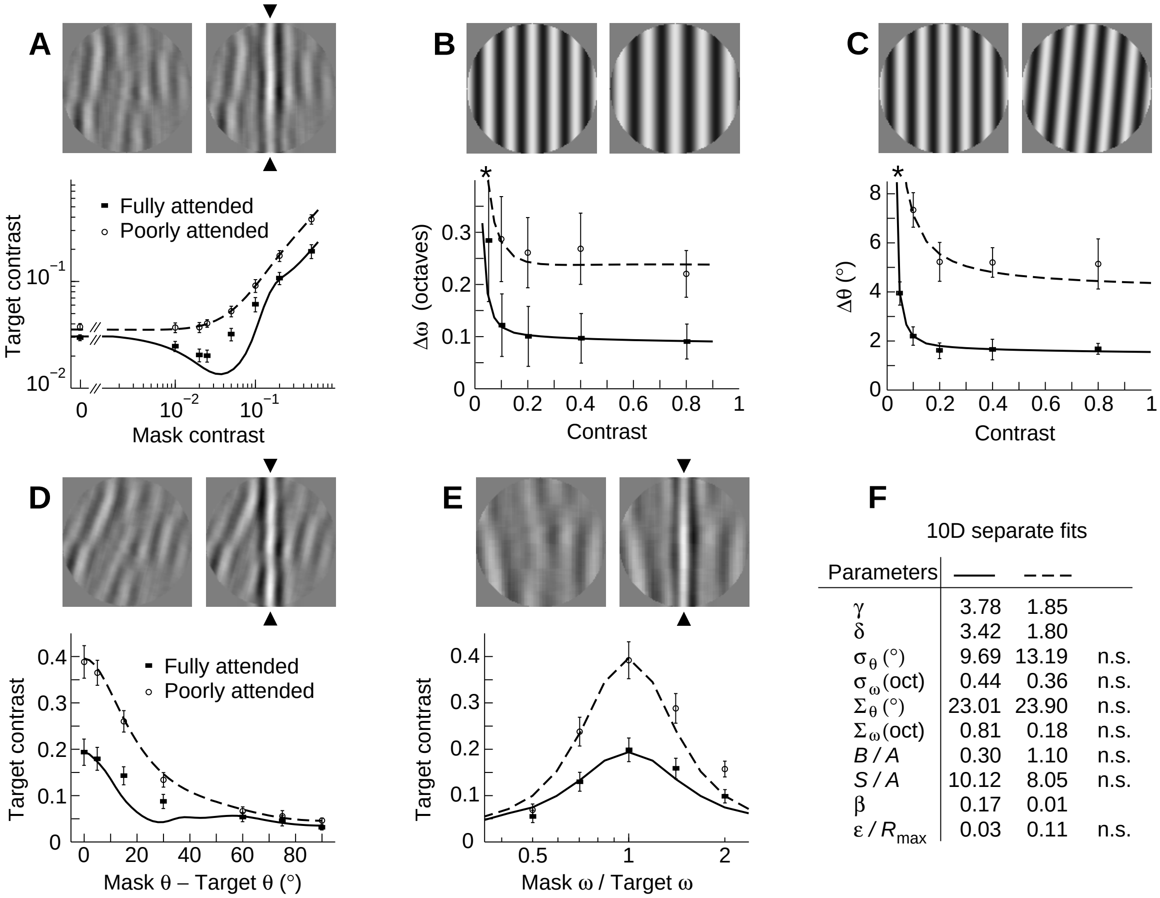
Next, we manually explore a number of simple parameter manipulations around the best-fit model to the fully-attended data. The manipulations investigated include separate modifications of filter gain, noise variance, filter tuning, inhibitory pool size and interaction strength. Of all these manipulations, only the alteration of the exponents γ and δ which govern the strength of interactions between filters yields a good prediction of the poorly-attended dataset.

In order to verify this finding in a systematic manner, we conduct two sets of simultaneous fits to the entire dataset (including both attentional conditions), using an extended model which makes two sets of predictions by allowing some of its parameters to change according to the attentional condition. In the first simultaneous fit, only γ and δ are permitted to assume different values in the two attentional conditions; all other parameters are constrained to identical values for both conditions. In the second simultaneous fit, all parameters except for γ and δ are allowed to change with attention.

8.3.3 Separate Fits to Either Attentional Condition

When we fit our model (10 free parameters: $\gamma, \delta, \sigma_\theta, \sigma_\omega, \Sigma_\theta, \Sigma_\omega, S, B, \beta, \epsilon$) separately to either single- or double-task data, we obtain good agreement between predicted and observed thresholds with physiologically plausible parameter values (solid curves in **Fig. 8.3**).

Note the realistic values for filter tuning, σ_θ and σ_λ: the half-widths at half maximum of model units are 12° to 15° and 0.42*oct* to 0.52*oct*, compared to approximately 20 ± 9° and 0.76 ± 0.30*oct* for neurons in monkey visual cortex (Geisler & Albrecht, 1997)). Note



Parameters	——	– – –	
γ	3.78	1.85	
δ	3.42	1.80	
σ_θ (°)	9.69	13.19	n.s.
σ_ω(oct)	0.44	0.36	n.s.
Σ_θ (°)	23.01	23.90	n.s.
Σ_ω(oct)	0.81	0.18	n.s.
B / A	0.30	1.10	n.s.
S / A	10.12	8.05	n.s.
β	0.17	0.01	
$\varepsilon / R_{\max}$	0.03	0.11	n.s.

Figure 8.3: Separate fits to the data. Solid curves represent fits computed separately for single- and double-task data (all 10 parameters are permitted to differ). Parameters marked “n.s.” were not significantly different when allowing the fitting error to increase by up to 5% above its best-fit value.

also that orientation and spatial frequency thresholds remain constant for contrast values above 20% **(Fig. 8.3.B,C)** and that the curves for full and poor attention appear displaced vertically rather than horizontally. This shows clearly that attention changes more than contrast gain, since a difference in gain of the linear filter stage would merely produce a horizontal displacement. The main discrepancy between model and data is that the model predicts a more pronounced dipper for contrast discrimination thresholds than is actually observed **(Fig. 8.3.A)**. Since the data may underestimate the dipper (Lee, 1999), this prediction may in fact be correct.

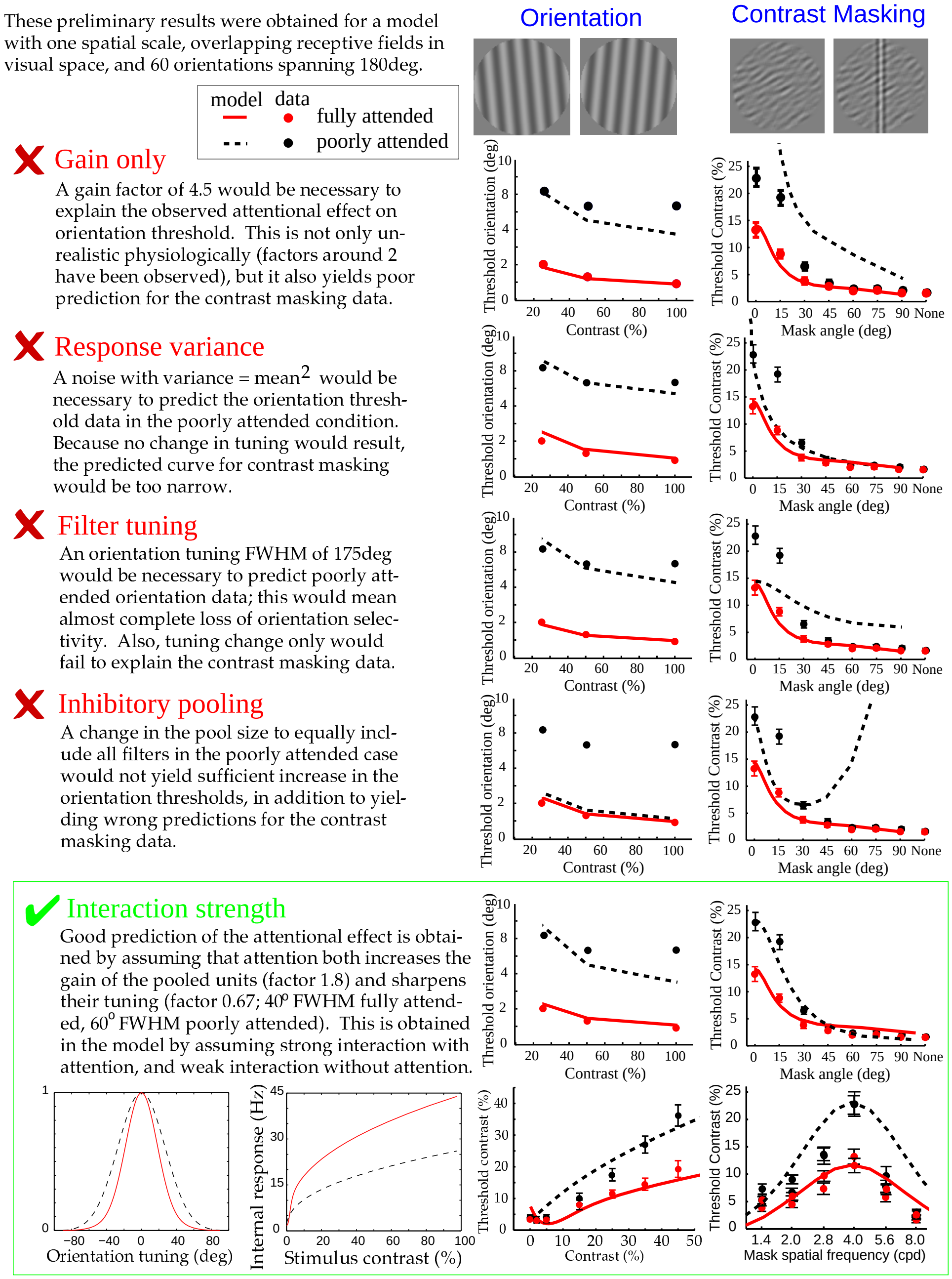
That a single set of parameter values accounts for all thresholds observed with either full or poor attention is not a matter of course. One might have expected that attending to stimulus orientation would affect visual processing differently than, say, attending to spatial frequency. In such case, different versions of our model would have been necessary to predict the different experiments. Instead, our results indicate that, under each attentional condition taken separately, a single model is capable of predicting all experiments. Consequently, not only do our results support our claims that the model represents a reasonably "unifying" account for basic spatial vision, but also they are consistent with the possibility that attention alters visual processing in the same way for all examined tasks.

8.3.4 Manual Alterations of the Model

In a preliminary step, we explore a number of manual alterations of the model **(Fig. 8.4)**. To this end, we started by fitting the model to the "fully attended" dataset, and subsequently manipulated some of the model's parameters in an attempt to predict the "poorly attended" data. For computational tractability of the manual interventions, these preliminary simulations only concerned contrast and orientation discriminations, and used a simpler version of our model, which only comprised one spatial scale. The manipulations investigated includes separate modifications of filter gain, noise variance, filter tuning, inhibitory pool size and interaction strength.

The parameters of the model were automatically adjusted to fit human psychophysical thresholds for contrast and orientation discrimination tasks, in the fully-attended condition. The model consisted of 60 orientations spanning 180° and one scale at 4 cycles per degree. The multidimensional simplex algorithm with simulated annealing overhead described in the previous Chapter was used to determine the best fit of the model to the fully-attended data. The free parameters adjusted during the automatic fits were: the noise level, the pooling exponents, the inhibitory pooling constant, the filter tuning widths, the inhibitory pool size, and the background firing rates (9 parameters in total, since there was no pooling across scales in this single-scale model).

As shown in **Fig. 8.4**, a change in both exponents γ and δ which govern the strength of interactions between the model units yielded a successful transition from predicting the



These preliminary results were obtained for a model with one spatial scale, overlapping receptive fields in visual space, and 60 orientations spanning 180deg.

✗ Gain only

A gain factor of 4.5 would be necessary to explain the observed attentional effect on orientation threshold. This is not only unrealistic physiologically (factors around 2 have been observed), but it also yields poor prediction for the contrast masking data.

✗ Response variance

A noise with variance = mean^2 would be necessary to predict the orientation threshold data in the poorly attended condition. Because no change in tuning would result, the predicted curve for contrast masking would be too narrow.

✗ Filter tuning

An orientation tuning FWHM of 175deg would be necessary to predict poorly attended orientation data; this would mean almost complete loss of orientation selectivity. Also, tuning change only would fail to explain the contrast masking data.

✗ Inhibitory pooling

A change in the pool size to equally include all filters in the poorly attended case would not yield sufficient increase in the orientation thresholds, in addition to yielding wrong predictions for the contrast masking data.

✓ Interaction strength

Good prediction of the attentional effect is obtained by assuming that attention both increases the gain of the pooled units (factor 1.8) and sharpens their tuning (factor 0.67; 40° FWHM fully attended, 60° FWHM poorly attended). This is obtained in the model by assuming strong interaction with attention, and weak interaction without attention.

Figure 8.4: Manual alterations of the model. In these preliminary experiments, a simple version of our detailed model (one scale, 60 orientations) was fit to the “fully attended” data, after which a number of parameters were manually altered in an attempt to obtain a prediction of the “poorly attended” data. Only the modification of interaction strength yielded satisfactory results.

fully-attended data to predicting the poorly-attended data.

The results from this computational effort suggest that attention does not simply change one particular property of early visual units (e.g., gain). Rather, we found that a combination of change in gain and tuning best explain attentional effects, as obtained through a modification of the strength of interactions between oriented units: Units interacted very strongly when attention is present, yielding a sigmoidal transducer and sharpening of the tuning curves. In the near absence of attention, interactions were much weaker, yielding a transducer closer to a simple power law, and no sharpening of the tuning curves.

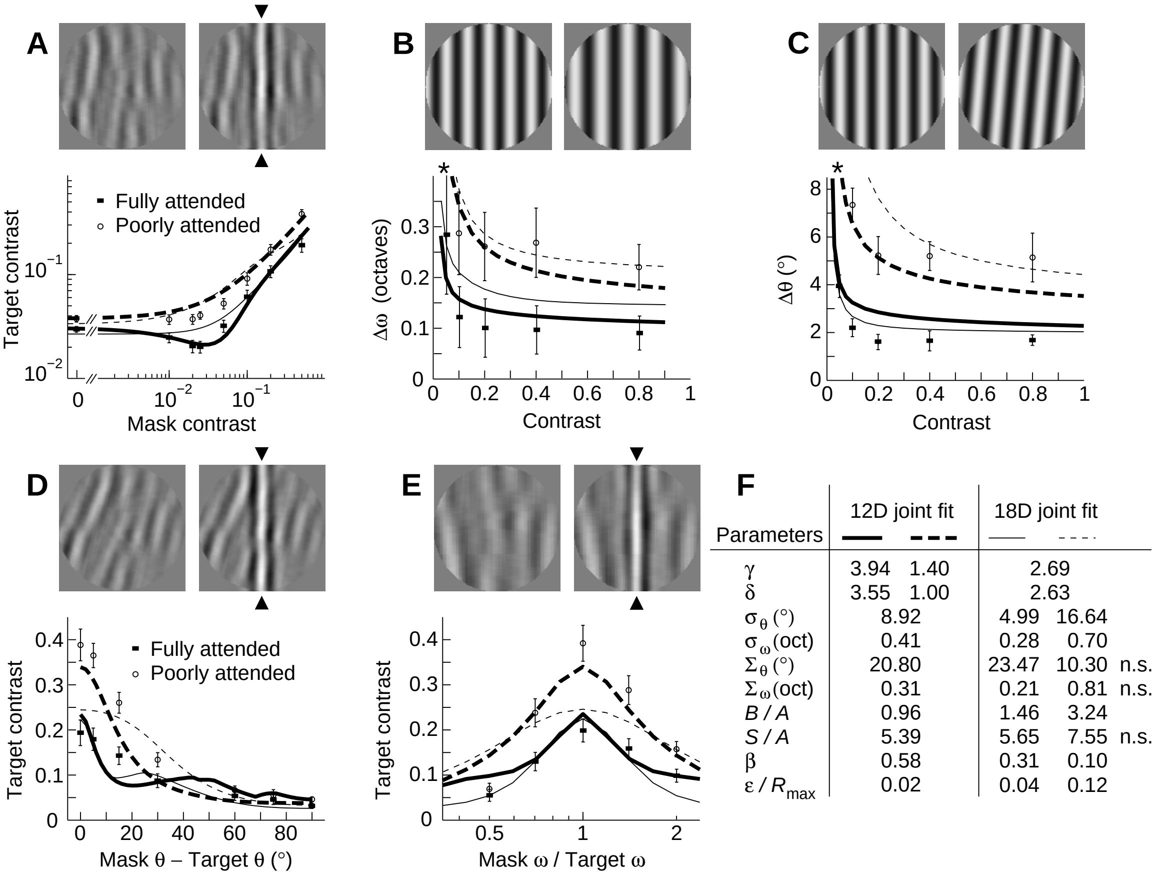
8.3.5 Simultaneous Automatic Fits to the Entire Dual-Task Dataset

Although there are several differences between the parameters obtained with full and poor attention, the change in the exponents of the power law, γ and δ, is especially noticeable. This was confirmed by a quantitative analysis of how well each parameter was constrained by the dataset (see Methods, Model fits and **Fig. 8.3.F**): All parameters except for γ, δ and β were not significantly different between both attentional conditions, when tolerating a small yet noticeable degradation of the quality of fit by less that 5%.

Additionally, the special role of these exponents was further underlined by the manual manipulations just described. We consequently decided to further study, this time in a systematic and unbiased manner, how well a single modification of the two exponents γ and δ could account for the attentional modulation of all observed thresholds.

To isolate the consequences of this change, we fit the model simultaneously to both single- and double-task data, while allowing *only* γ, δ to take different values depending on attention (12 free parameters). In other words, γ, δ take two values while all other parameters take a single value. This assumes that attention alters *only* the strength of the interaction between filters. Even with this restrictive assumption, we obtain acceptable fits with physiologically plausible parameter values ("12D joint fits," solid curves in **Fig. 8.5.A-E**, two leftmost columns in **Fig. 8.5.F**).

Contrariwise, when we allow *all* parameters *except* γ, δ to take different values depending on attention (18 free parameters), there are no acceptable fits with plausible parameter values ("18D joint fits," dashed curves in **Fig. 8.5.A-E**, two rightmost columns in **Fig. 8.5.F**). The best possible fit turns out to be rather poor, and predicts neither the dipper in the contrast discrimination thresholds **(Fig. 8.5.A)** nor the maximal extent of contrast masking **(Fig. 8.5.D,E)**. Furthermore, even this poor fit requires unrealistic parameter values **(Fig. 8.5.F)**: Attention would have to alter the tuning widths of visual filters (σ_θ from 17° to 5° and σ_λ from $0.7oct$ to $0.3oct$) and, in addition, would have to "turn on" inhibitory pooling in the orientation dimension (Σ_θ from $0.6\sigma_\theta$ to $5\sigma_\theta$), without changing it in the spatial frequency dimension (Σ_λ from $1.2\sigma_\lambda$ to $1.3\sigma_\lambda$). Such drastic changes in cortical interactions do not appear to be plausible.



Parameters	12D joint fit —	- - -	18D joint fit —	- - -	
γ	3.94	1.40	2.69		
δ	3.55	1.00	2.63		
σ_θ (°)	8.92		4.99	16.64	
σ_ω(oct)	0.41		0.28	0.70	
Σ_θ (°)	20.80		23.47	10.30	n.s.
Σ_ω(oct)	0.31		0.21	0.81	n.s.
B / A	0.96		1.46	3.24	
S / A	5.39		5.65	7.55	n.s.
β	0.58		0.31	0.10	
$\varepsilon / R_{\max}$	0.02		0.04	0.12	

Figure 8.5: Joint fits to the data for both attentional conditions. Predicted thresholds when attention changes some model parameters but not others. The solid curves represent a simultaneous fit to both single- and double-task data, in which only the exponents γ and δ take different values depending on attention (12 free parameters). Observed and predicted thresholds agree reasonably well, and parameter values are physiologically plausible (two leftmost columns in **F**). The dashed curves represent the optimal joint fit when all parameters *except* the exponents γ, δ take different values depending on attention (18 free parameters). Neither the dipper **(A)** nor the maximal extent of contrast masking **(D,E)** are predicted, and parameter values are unrealistic (two rightmost columns in **F**). Parameters marked (n.s.) were not significantly different when allowing the fitting error to increase by up to 5% above its best-fit value.

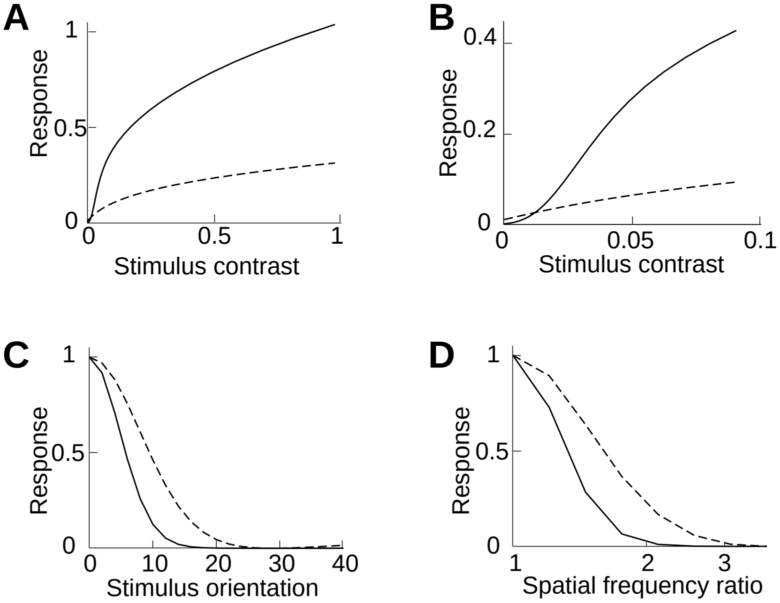


Figure 8.6: Effect of attention on early visual processing. Predictions based on 12-dimensional joint fit in **Fig. 8.5.F**. Attention increases the contrast gain (3.3-fold, **A**), causes the contrast response to assume sigmoidal shape at low contrast **(B)**, sharpens orientation tuning (by 40%, **C**) and spatial-frequency tuning (by 30%, **D**). To the extent that the visual filters of our model reflect individual neurons in visual cortex, this predicts that attention both increases the gain and sharpens the tuning of such neurons.

In short, the effect of attention can be reproduced in a quantitative manner by selectively increasing the exponents of the power law, γ and δ, but not by changing other parts of the model. For example, changes of noise parameters, β, ϵ, tuning of linear filters, σ_θ, σ_ω, or size of the inhibitory pool, Σ_θ, Σ_ω, cannot account for the observed effects of attention. The question as to *why* larger values of γ and δ should account for attention is taken up in the discussion.

8.4 Discussion

The observations of Lee *et al.* tightly constrain any effect attention may have on the visual filters and/or the interactions amongst filters that are thought to underlie basic pattern vision. Comparison with a computational model shows that the observed effects of attention are consistent with stronger interactions amongst filters, but not with a change in noise parameters without change in interactions, as is sometimes thought (Bonnel & Miller, 1994; Lu & Dosher, 1998). Essentially, the effects of attention on different thresholds are too disparate to be accommodated by a single change in noise parameters.

In the framework of our model, the strength of the interactions amongst filters is controlled by the exponents of a power law, γ and δ. The immediate reasons why larger exponents (*i.e.*, values of 3.5 instead of 1.5) account for the observed effects of attention are as follows: Higher exponents accentuate the sigmoidal shape of the contrast response function at small contrasts **(Fig. 8.6.A)**, which explains the enhanced dipper of the contrast discrimination curve. The altered contrast response also reduces background activity and, thus, lowers contrast detection thresholds. At larger contrast, higher exponents entail an 3.3-fold increase in contrast gain **(Fig. 8.6.B)**, which accounts for lower contrast discrimination and contrast masking thresholds. Additionally, higher exponents sharpen the tuning for orientation (by 40%, **Fig. 8.6.C**) and spatial frequency (by 30%, **Fig. 8.6.D**), which sharply reduces thresholds for discriminating orientation and spatial frequency.

The more fundamental reason is, however, that larger exponents activate what is best described as a *winner-take-all competition amongst visual filters.* **Fig. 8.7** shows how attention (larger exponents) shifts the distribution of responses across the population of filters. Attention accentuates existing differences between filter responses, boosting filters that respond relatively well to a given stimulus, while suppressing filters that respond relatively poorly. Another way of putting it is that the distribution of responses is far narrower with full than with poor attention. This explains the perceptual advantage conferred by attention: Attention enhances the sensory representation by restricting responses to the filters tuned best to the stimulus at hand. To the extent that visual filters can be identified with individual neurons in visual cortex, our model thus predicts that attention changes both the gain and tuning of such neurons.

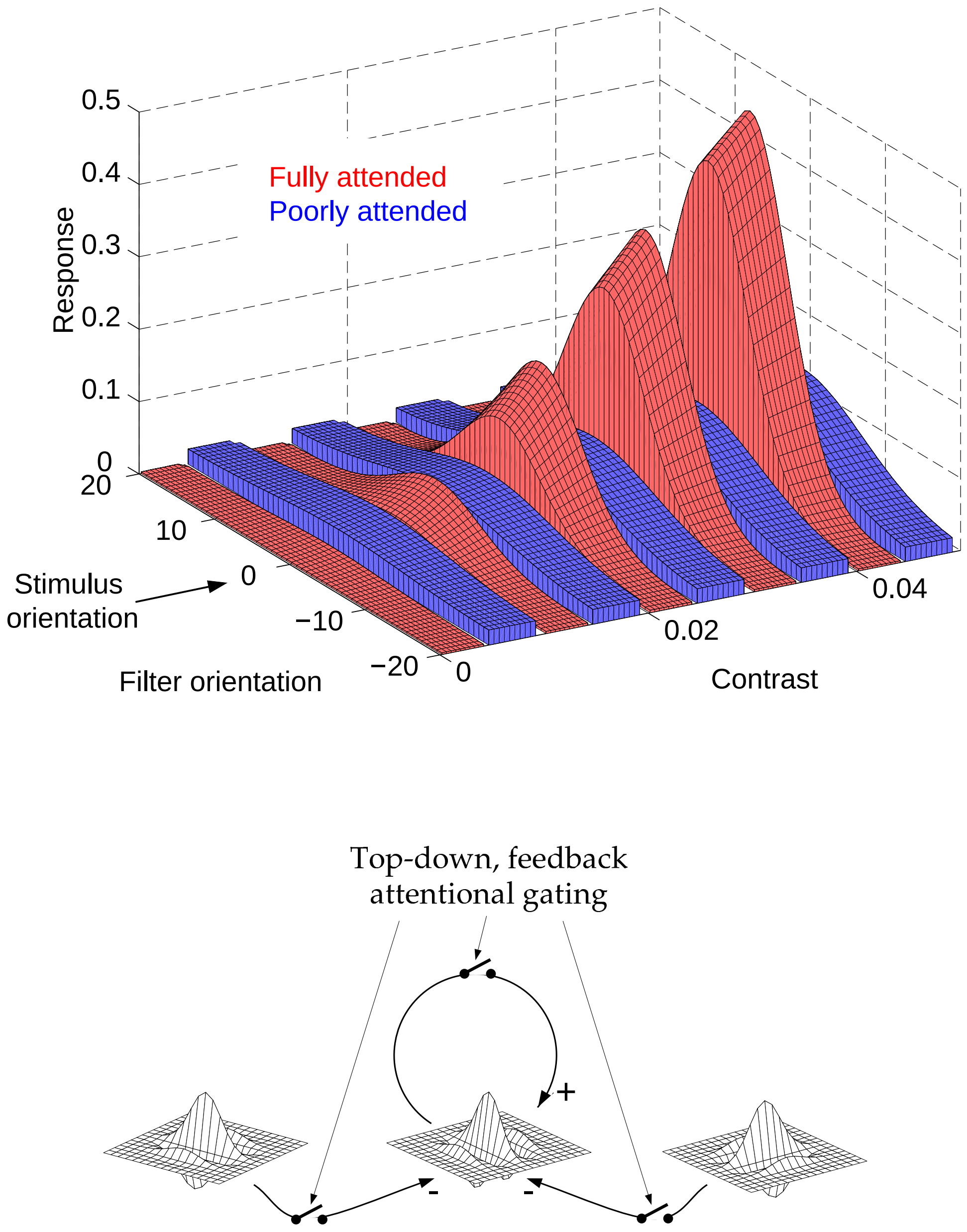


Figure 8.7: Attentional change in the response distribution. The top plot shows predictions based on the 12-dimensional joint fit in **Fig. 8.5.F**. Responses $R_{\theta,\omega}$ of filters tuned to orientations between $-20°$ to $+20°$ to a grating stimulus of orientation $0°$ and contrast between 0 to 0.05 (threshold regime). Responses to fully and poorly attended stimuli are represented by the red and blue surfaces, respectively (shown interleaved for clarity). By strengthening a winner-take-all competition amongst visual filters, attention restricts responses to the filters tuned best to the stimulus at hand. The bottom plot shows a simple functional schematic of the top-down modulatory action of attention as inferred from our model.

Our model is consistent with recent findings in the visual cortex of humans and monkeys. Attentional changes in neuronal activity have been reported in several early visual cortical areas, including areas V1, V2, V4, and MT/MST (Moran & Desimone, 1985; Motter, 1993; Treue & Maunsell, 1996; Luck *et al.*, 1997; Roelfsema *et al.*, 1998; Gandhi *et al.*, 1998; Brefczynski & DeYoe, 1998). Furthermore, the notion that attention modulates a local competition in visual cortex has been proposed independently on the basis of theoretical (Niebur & Koch, 1994; Tsotsos *et al.*, 1995) and single-unit studies (Desimone, 1998). In the macaque, attentional modulation of responses in the visual cortex is weak or absent if only a single stimulus is present in the receptive field, suggesting that attention modulates interactions between neurons with overlapping receptive fields (Moran & Desimone, 1985; Luck *et al.*, 1997).

Our model is consistent with reports that attention increases contrast gain in areas V2 and V4 of the macaque (McAdams & Maunsell, 1996b; Reynolds *et al.*, 1997). Whether attention sharpens the orientation tuning of visual cortical neurons remains unclear from these experiments (Spitzer *et al.*, 1988; McAdams & Maunsell, 1996b). Indeed, our results suggest that an approximately 35% in sharpening of the tuning curves accompanies the predicted 3.3-fold increase in gain. A very small sharpening of the tuning curves, of the order of 1% to 1.5%, would consequently be expected to correspond to the 10% to 15% gain modulation observed in single-unit studies; while such sharpening has not been observed (Treue & Trujillo, 1999), it may just be for technical reasons related to experimental precision in the measurements. Unfortunately, while dual-task experiments appear to yield much stronger attentional modulation than the simple "attend to location" paradigms typically used in electrophysiological studies, it seems impractical to attempt to conduct such experiments (which already are quite difficult for human observers) with awake monkeys.

We suggest that one interesting manner in which an experimental paradigm suitable for monkey experiments could be designed is through the use of functional imaging techniques in humans. Indeed, using BOLD-contrast imaging, we have been able to see (and others before us (Gandhi *et al.*, 1998)) large attentional effect in areas V1 and V2 when performing simple visual discrimination tasks. In addition to providing direct support for our general result that attention increases, through top-down feedback, activity in these early cortical areas **(Fig. 8.7)**, variations around these simple imaging experiments could be used to calibrate and optimize the strength of attentional modulation obtained for given experimental paradigms. Indeed, a 4-minute scan is sufficient in our experience to derive quantitative measures, to some degree, of how strong an attentional modulation can be expected for a given set of stimuli and task instructions **(Fig. 8.8)**.

It is also worth noting that our model predicts that fully focussed attention sharpens orientation tuning in the parts of visual cortex that mediate basic pattern vision (presumably areas V1 and/or V2). Increased competition in area V4 or MT could, however, result in

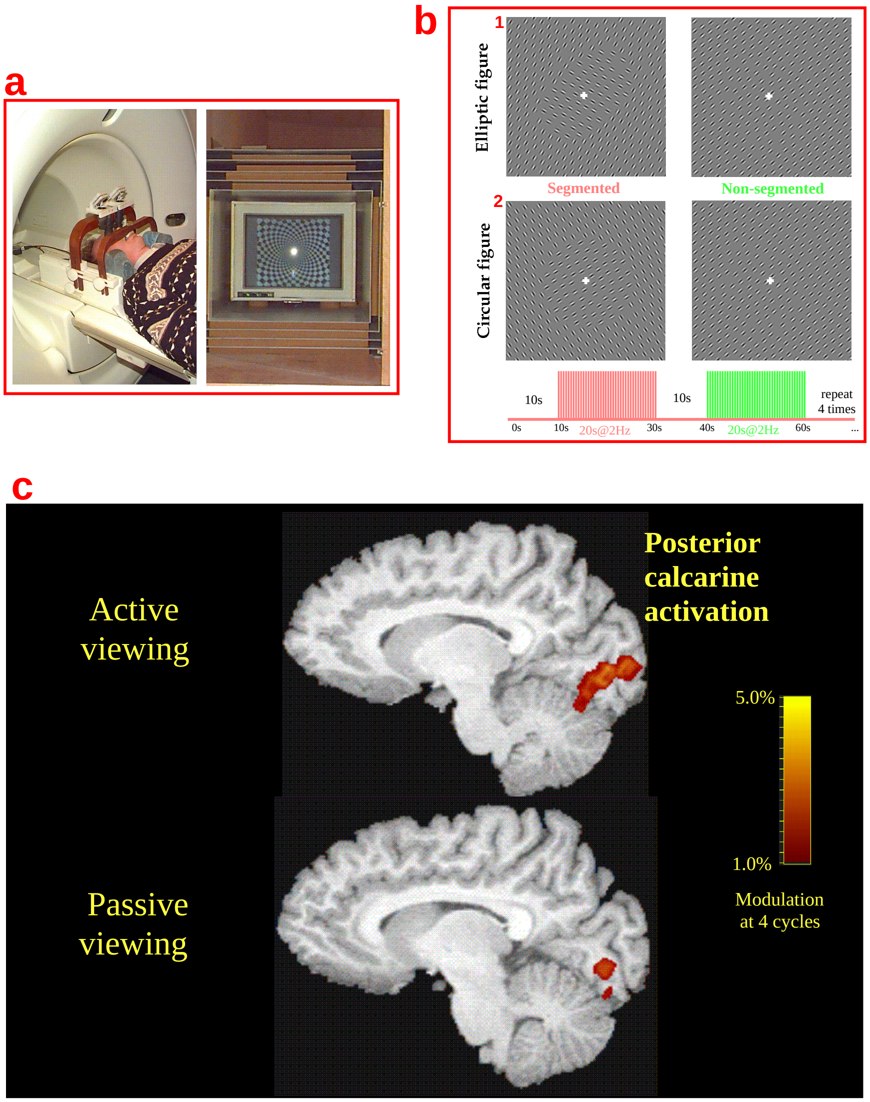


Figure 8.8: Using fMRI to prototype and optimize attentional modulation experiments. This figure shows strong attentional modulation in areas V1 and V2 during an experiment conducted by Drs. Jochen Braun, Linda Chang and Thomas Ernst at the Harbor-UCLA Medical Center (Torrance, CA), and is part of a broader study of visual segmentation by these investigators. **(a)** I am lying in the scanner and viewing, through a system of binoculars and mirrors, stimuli presented on a CRT screen shielded from the ambient magnetic field by several layers of μ-metal. **(b)** The experiment uses a blocked paradigm in which observers discriminate between the stimuli shown. The only important experimental characteristic for the purpose of this illustration is that observers can either passively view the stimuli (implying a state of low and diffuse attention), or have to actively perform a shape discrimination task (implying strongly focused attention). **(c)** Activation along the posterior calcarine sulcus is much stronger when observers have to engage attention onto the task compared to when they passively view the stimuli. Repeating these experiments for a variety of stimulus configurations allowed the investigators to optimize the strength of the observed effect; here we suggest that similar methods could be used to rapidly design and optimize experiments to be eventually performed in monkeys.

sharpening for a more complicated stimulus dimension.

Finally, we do not wish to claim that attention is restricted to local interactions at one particular level of visual cortex. More than likely, attention has additional effects on long-range interactions at the same level and indeed at all levels of visual cortex. Nevertheless, our results show that the activation of a winner-take-all competition amongst overlapping visual filters explains many basic perceptual consequences of attention.

8.5 Outlook

In this Chapter, we have seen that the simple model of basic pattern vision presented in the previous Chapter is capable of predicting attentional modulation of a variety of psychophysical thresholds in a simple and task-independent manner. This result was obtained by first deriving a set of modeling constraints from the data by Lee *et al.*, using a simplified model in which different filters do not interact. We then used our detailed model to carry out fully detailed quantitative simulation of the entire dataset.

Not only was the model able to reproduce all thresholds separately for either attentional condition, but we also found that a simple change in the strength of interactions among visual filters was sufficient to simultaneously transition all the model's predictions from poorly-attended to fully attended thresholds. This allowed us to advance the very simple computational hypothesis that attention activates a winner-take-all competition among early visual filters.

In the next Part of this thesis, we present a general discussion which relates to both aspects of our work, the bottom-up and top-down functional expression of visual attention.

Part IV

General Outlook

This last part of the present thesis summarizes our findings and their meanings in terms of computation in neurobiological systems.

8.6 Summary of Our Results on Modeling Bottom-Up Attention

In **Part II** of this thesis, we have proposed a detailed model for the control of visual attention based uniquely on bottom-up, image-driven cues.

We have shown that the problem of combining several multiscale feature maps into a unique saliency map can be best solved, in a generic manner, by endowing the feature maps with simple spatially competitive dynamics. We have then shown that using a simple winner-take-all network and inhibition-of-return mechanism was an efficient neuronal implementation for a simple attention focusing mechanism.

Despite the fairly crude implementation of our model, we have found that it is capable of strong performance at detecting salient objects in complex natural scenes. These results were obtained using images which always contained some amount of noise, or degradations due to their storage format, and had not been carefully optimized for the model. Indeed, most of our results are based on databases of images which were submitted to us by external collaborators. In addition, most scenes studied were complex, with typically a large number of objects, sharp contrasts in illumination conditions and color appearances, shadows and occlusions. This contrasts with many computer vision models which are too often restricted to grayscale images containing a small set of simple shapes on a uniform background.

Although the reproduction by our model of simple visual search tasks could have been expected, it was not a matter of course when using the noisy stimuli which we described. With our stimuli which contained speckle noise, spatial jitter noise, and array element orientation noise, the basic difference between pop-out and conjunctive search was not always obtained when using a “naive” model, in which all feature maps are simply summed into the saliency map. One of our main findings was the realization that including some non-linear dynamical computation already at the earliest levels of processing was key to the performance of the model, especially when comparing it to humans.

A second important finding was that our model performs close to, or even sometimes better than, humans do on similar tasks. This result was achieved even though our model implements one of the most severe information processing bottlenecks possible: In our model, the attentional focus is entirely controlled by activity in a scalar map, the saliency map, which is 16 times smaller in x and y than the input image. Despite such drastic reduction of information, **by a factor 768** between the trichromatic retinal input and the saliency map, our model often behaves as if it had a clear knowledge of all the details of the scene. This remarkable result represents strong evidence for the computational validity of Koch and Ullman's (1985) original idea that a single, explicitly represented saliency map could efficiently guide attention. This very concise and explicit representation is in sharp contrast with a number of theories of attention, in which attention is an implicit, emergent property to which all visual areas directly contribute.

8.7 Summary of Our Results on Modeling Spatial Vision Thresholds

In **Part III** of this thesis, we have proposed a very small model of just one hypercolumn in primary visual cortex. Despite its small size, with an input layer comprising just 60 spatially bandpass filters, this model was able to produce a simultaneous quantitative account of 34 human psychophysical thresholds, for five different classes of pattern discrimination experiments.

Such result was obtained in part through our original use of a generic theory for ideal observer discrimination in any type of psychophysical task. Using such framework allowed us to overcome the problem, previously reported as unsolvable, of simultaneously predicting contrast and orientation discrimination thresholds using a single set of model parameters. This framework was directly developed following existing models which had used similar concepts in simplified situations. Here, we have been able to extend it to a realistic situation, in which any discrimination task may be considered, and in which a sparse population of analyzing filters may be used.

Also of importance was our insistence in carrying out fully detailed numerical simulations of the full non-linear model, and in avoiding linear or constant approximations each time they were not fully justified and accurate. Although this resulted in our model taking several days of computation time to converge towards a full prediction, and required the development of specific computer process dispatch and management tools, we believe that it also allowed our model to predict some of the more detailed features of our experimental dataset. Indeed, we have described in several instances throughout this thesis both that simplified versions of our model could only yield fairly approximate or sometimes wrong predictions, and that important results such as the capability of simultaneously predicting contrast and orientation thresholds was due to an apparently insignificant detail, namely a slight broadening of the tails of our units' tuning curves with contrast.

Our results are also remarkable in the fact that, in addition to its small size, our model is very coarse in its implementation. Indeed, it does not include any refinement such as synchronization among neurons, spike timing, or even any spiking for that matter. Although our model only represents a functional account for the putative neuronal circuits responsible for basic pattern vision, we have found that when all model parameters were left entirely free to be determined by our psychophysical data, they converged towards plausible neuronal values. This strong result indicates that our formulation of the model, using simple linear filtering followed by non-linear interactions in the form of power law and divisive inhibition among neurons with similar tuning properties, may well represent a unifying understanding of the basic functional principles underlying perception of simple patterns.

In part, our successful account for a wide variety of thresholds is due to our acquisition of a coherent dataset in several human observers. By using localized pattern stimuli with always the same shape, and by presenting them in the near periphery such that they would presumably only excite one hypercolumn in primary visual cortex, our dataset represents one of the most comprehensive accounts for basic spatial vision.

8.8 Summary of Our Results on Modeling Top-Down Attentional Modulation

When we further applied our model of basic pattern vision to the dual-task dataset of Lee *at al.*, a first remarkable result was that this model, which had originally been designed to account for thresholds under full attention, presented no difficulty in simultaneously predicting all thresholds under poor attention as well. This strong result reinforces the idea that our model may well represent a unified functional account for the earliest cortical stages of visual processing in primates.

In addition, not only were the internal characteristics of the model fit to the poorly attended dataset reasonable, as they had been for the fully attended dataset, but they directly pointed us towards a simple computational explanation for the observed attentional modulation. Indeed, although the separate fits to the fully and poorly attended thresholds had been conducted independently, and involved, in each case, random starting parameter values and no other constraint onto the model than dictated by the fitting error to the data, we found that only three model parameter differed significantly between both attentional conditions, namely one noise parameter and two parameters regulating the strength of interactions among model units.

By carrying joint fits of a dual model to the entire dual-task dataset, we further showed that the attentional modulation in noise, alone or combined with any additional modulation in any parameter except for those regulating the interactions among units, could not explain the observed psychophysical expression of top-down attentional feedback. The initially different value of the noise parameter observed in the separate fits hence appeared not to be critical to the results. A final joint fit in which attention could only alter the values of the two exponents γ and δ which expressed the strength of interactions yielded a successful, quantitative and simultaneous account for all dual-task observations.

These results led us to propose, on the basis of the increase in gain and sharpening of the tuning curves resulting from the strengthening of interactions with attention, that attention activates a winner-take-all competition among early visual filters. On the one hand, one remarkable aspect of this theory is that it represents an extremely simple, task-independent and prototypical account for attentional modulation, and is entirely expressed at the earliest stages of cortical visual processing. This contrasts with, although it does not necessarily preclude or contradict, former theories which envisioned attentional modulation as the result of a very complicated, task-aware and thus cognitively-controlled, process.

On the other hand, our theory appears to yield a much richer variety of predictions than previously proposed much simpler theories based on electrophysiological experiments, which regard attentional feedback as a multiplicative gain modulation. While we have seen that, if we scale the attentional effects observed in our psychophysical experiments down to the level of the effects observed in physiology, our theory would also reduce to a small gain modulation with negligible sharpening of tuning, we have also provided strong experimental and modeling evidence that such gain modulation cannot simultaneously explain the modulation of all of our observed thresholds. As electrophysiological experiments typically only concern a few of our 34 datapoints, we conclude at present that more single-unit studies, if possible demonstrating stronger attentional modulation, will in the future provide a critical test for our theory.

8.9 Computation in Neural Systems

As it has probably become clear to the reader, the common emerging theme in the various parts of this thesis is that non-linear computation at the earliest levels of visual processing are the main contributors to the observed performance of the models which we developped.

Our models, both the one concerned with bottom-up and the one concerned with top-down attention, use linear filtering stages as front-ends. One of the main conclusions from our bottom-up model, however, was that combining the results of these linear pre-processing stages into the saliency map yielded very poor predictions; this prompted us to include more computation into the very first levels of processing, by implementing a non-linear spatial competition for salience in every feature map.

Similarly, one of the main conclusions from our model of early pattern vision and its modulation by top-down attention was that the predictive power of the model, and its ability to explain the observed effect of attention, entirely lied in the non-linear interactions among visual filters. Not only are these interactions the component of the model which allowed us to produce a simultaneous fit of all of our data, but they also are the component which is modulated by attention.

To conclude this thesis, I consequently would like to leave the reader with the simple idea that the more computation we have added to the earliest stages of our models, the better and closer to human vision have the predictions of these models become. This convinced me that early stages of visual processing are far from being adequately modeled by simple linear filters and convolution operators. While such classical approximation may reasonably well capture the amount of processing achieved by visual neurons when stimulated with a bright dot or a moving bar on a uniformly dark background, we have seen that this is no more the case as soon as several stimuli are presented in the visual environment, at which point both long-range and short-range non-linear connections become predominant factors in shaping the neuronal responses.

References

Ahmed, B., Allison, J. D., Douglas, R. J., & Martin, K. A. 1997. An intracellular study of the contrast-dependence of neuronal activity in cat visual cortex. *Cereb Cortex*, **7**(6), 559–70.

Albrecht, D. G., & Geisler, W. S. 1991. Motion selectivity and the contrast-response function of simple cells in the visual cortex. *Vis Neurosci*, **7**(6), 531–46.

Albrecht, D. G., & Hamilton, D. B. 1982. Striate cortex of monkey and cat: contrast response function. *J Neurophysiol*, **48**(1), 217–37.

Allison, J. D., Casagrande, V. A., & Bonds, A. B. 1995. The influence of input from the lower cortical layers on the orientation tuning of upper layer V1 cells in a primate. *Vis Neurosci*, **12**(2), 309–20.

Amir, Y., Harel, M., & Malach, R. 1993. Cortical hierarchy reflected in the organization of intrinsic connections in macaque monkey visual cortex. *J Comp Neurol*, **334**(1), 19–46.

Andersen, R. A. 1997. Multimodal integration for the representation of space in the posterior parietal cortex. *Philos Trans R Soc Lond B Biol Sci*, **352**(1360), 1421–8.

Andersen, R. A., Bracewell, R. M., Barash, S., Gnadt, J. W., & Fogassi, L. 1990. Eye position effects on visual, memory, and saccade-related activity in areas LIP and 7a of macaque. *J Neurosci*, **10**(4), 1176–96.

Ben-Yishai, R., Bar-Or, R. L., & Sompolinsky, H. 1995. Theory of orientation tuning in visual cortex. *Proc Natl Acad Sci U S A*, **92**(9), 3844–8.

Bergen, J. R., & Julesz, B. 1983. Parallel versus serial processing in rapid pattern discrimination. *Nature*, **303**(5919), 696–8.

Beymer, D., & Poggio, T. 1996. Image Representations for Visual Learning. *Science*, **272**(5270), 1905–9.

Bijl, P., Kooi, F. K., & van Dorresteijn, M. 1997. *Visual search performance for realistic target imagery from the DISSTAF field trials.* Soesterberg, The Netherlands: TNO Human Factors Research Institute.

Blasdel, G. G., & Salama, G. 1986. Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. *Nature*, **321**(6070), 579–85.

Blasdel, G. G., Lund, J. S., & Fitzpatrick, D. 1985. Intrinsic connections of macaque striate cortex: axonal projections of cells outside lamina 4C. *J Neurosci*, **5**(12), 3350–69.

Bolz, J., Gilbert, C. D., & Wiesel, T. N. 1989. Pharmacological analysis of cortical circuitry. *Trends Neurosci*, **12**(8), 292–6.

Bonnel, A. M., & Miller, J. 1994. Attentional effects on concurrent psychophysical discriminations: investigations of a sample-size model. *Percept Psychophys*, **55**(2), 162–79.

Bonnel, A. M., Possamai, C. A., & Schmitt, M. 1987. Early modulation of visual input: a study of attentional strategies. *Q J Exp Psychol [A]*, **39**(4), 757–76.

Bonnel, A. M., Stein, J. F., & Bertucci, P. 1992. Does attention modulate the perception of luminance changes? *Q J Exp Psychol [A]*, **44**(4), 601–26.

Borgefors, G. 1991. Distance transformations in digital images. *Page 301 of: CVGIP: Image Understanding*, vol. 54.

Bowen, R. W., & Wilson, H. R. 1994. A two-process analysis of pattern masking. *Vision Res*, **34**(5), 645–57.

Bowne, S. F. 1990. Contrast discrimination cannot explain spatial frequency, orientation or temporal frequency discrimination. *Vision Res*, **30**(3), 449–61.

Bradley, A., & Skottun, B. C. 1984. The effects of large orientation and spatial frequency differences on spatial discriminations. *Vision Res*, **24**(12), 1889–96.

Bradley, A., Skottun, B. C., Ohzawa, I., Sclar, G., & Freeman, R. D. 1985. Neurophysiological evaluation of the differential response model for orientation and spatial-frequency discrimination. *J Opt Soc Am [A]*, **2**(9), 1607–10.

Braun, J. 1994. Visual search among items of different salience: removal of visual attention mimics a lesion in extrastriate area V4. *J Neurosci*, **14**(2), 554–67.

Braun, J. 1998a. Vision and attention: the role of training. *Nature*, **393**, 424–425.

Braun, J. 1998b. Vision and attention: the role of training [letter; comment]. *Nature*, **393**(6684), 424–5. Comment on: Nature 1997 Jun 19;387(6635):805-7.

Braun, J., & Julesz, B. 1998. Withdrawing attention at little or no cost: detection and discrimination tasks. *Percept Psychophys*, **60**(1), 1–23.

Braun, J., & Sagi, D. 1990. Vision outside the focus of attention. *Percept Psychophys*, **48**(1), 45–58.

Brefczynski, J. A., & DeYoe, E. A. 1998. A physiological correlate of the spotlight of visual attention. *Soc Neurosci Abstr*, **24**, 493.7.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. 1992. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J Neurosci*, **12**, 4745–4765.

Burbeck, C. A., & Regan, D. 1983. Independence of orientation and size in spatial discriminations. *J Opt Soc Am*, **73**(12), 1691–4.

Burt, P.J., & Adelson, E.H. 1983. The Laplacian Pyramid as a Compact Image Code. *IEEE Trans. on Communications*, **31**, 532–540.

Cannon, M. W., & Fullenkamp, S. C. 1991. Spatial interactions in apparent contrast: inhibitory effects among grating patterns of different spatial frequencies, spatial positions and orientations. *Vision Res*, **31**(11), 1985–98.

Cannon, M. W., & Fullenkamp, S. C. 1996. A model for inhibitory lateral interaction effects in perceived contrast. *Vision Res*, **36**(8), 1115–25.

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

Carandini, M., & Ringach, D. L. 1997. Predictions of a recurrent model of orientation selectivity. *Vision Res*, **37**(21), 3061–71.

Carandini, M., Heeger, D. J., & Movshon, J. A. 1997. Linearity and normalization in simple cells of the macaque primary visual cortex. *J Neurosci*, **17**(21), 8621–44.

Carandini, M., Movshon, J. A., & Ferster, D. 1998. Pattern adaptation and cross-orientation interactions in the primary visual cortex. *Neuropharmacology*, **37**(4-5), 501–11.

Cheal, M, & Lyon, D R. 1994. Attention in visual search: multiple search classes. *Quart J Exp Psychol A*, **47**, 49–69.

Chelazzi, L, & Desimone, R. 1994. Responses of V4 neurons during visual search. *Soc Neurosci Abstr*, **20**, 1054.

Colby, C. L., & Goldberg, M. E. 1999. Space and attention in parietal cortex. *Annu Rev Neurosci*, **22**, 319–49.

Connor, C. E., Gallant, J. L., Preddie, D. C., & Essen, D. C. Van. 1996. Responses in area V4 depend on the spatial relationship between stimulus and attention. *J Neurophysiol*, **75**(3), 1306–8.

Connor, C. E., Preddie, D. C., Gallant, J. L., & Essen, D. C. Van. 1997. Spatial attention effects in macaque area V4. *J Neurosci*, **17**(9), 3201–14.

Corbetta, M. 1998. Frontoparietal cortical networks for directing attention and the eye to visual locations: identical, independent, or overlapping neural systems? *Proc Natl Acad Sci U S A*, **95**(3), 831–8.

Cover, T. M., & Thomas, J. A. 1991. *Elements of Information Theory*. New York, NY: Wiley.

Cowey, A., & Rolls, E. T. 1974. Human cortical magnification factor and its relation to visual acuity. *Exp Brain Res*, **21**(5), 447–454.

Crick, F., & Koch, C. 1998. Consciousness and neuroscience. *Cerebral Cortex*, **8**, 97–107.

Crook, J. M., Kisvarday, Z. F., & Eysel, U. T. 1997. GABA-induced inactivation of functionally characterized sites in cat striate cortex: effects on orientation tuning and direction selectivity. *Vis Neurosci*, **14**(1), 141–58.

Dai, H. 1994. Signal-frequency uncertainty in spectral-shape discrimination: psychometric functions. *J Acoust Soc Am*, **96**(3), 1388–96.

Dai, H., Nguyen, Q., & Green, D. M. 1996. Decision rules of listeners in spectral-shape discrimination with or without signal-frequency uncertainty. *J Acoust Soc Am*, **99**(4 Pt 1), 2298–306.

Das, A., & Gilbert, C. D. 1999. Topography of contextual modulations mediated by short-range interactions in primary visual cortex [see comments]. *Nature*, **399**(6737), 655–61. Comment in: Nature 1999 Jun 17;399(6737):641, 643-4.

DeAngelis, G. C., Robson, J. G., Ohzawa, I., & Freeman, R. D. 1992. Organization of suppression in receptive fields of neurons in cat visual cortex. *J Neurophysiol*, **68**(1), 144–63.

Deneve, S., Latham, P. E., & Pouget, A. 1999. Reading population codes: a neural implementation of ideal observers. *Nat Neurosci*, **2**(8), 740–5.

Desimone, R. 1998. Visual attention mediated by biased competition in extrastriate visual cortex. *Phil Trans R Soc Lond B*, **353**, 1245–1255.

Desimone, R., & Duncan, J. 1995. Neural mechanisms of selective visual attention. *Annu Rev Neurosci*, **18**(Rev), 193–222.

DeValois, R. L., & DeValois, K. K. 1988. *Spatial Vision*. New York: Oxford University Press.

DeValois, R. L., Albrecht, D. G., & Thorell, L. G. 1982. Spatial-frequency selectivity of cells in macaque visual cortex. *Vision Research*, **22**, 545–559.

Dosher, B. A., & Lu, Z. L. 1997. Attention to location mediated by internal noise-reduction. *Inv Oph Vis Sci*, **38**(4 Part 2), 3205.

Douglas, R., Koch, C., Mahowald, M., Martin, K. A. C., & Suarez, H. H. 1995a. Recurrent excitation in neocortical circuits. *Science*, **269**, 981–985.

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

Douglas, R J, Martin, K A C, & Whitteridge, D. 1991. An intracellular analysis of the visual responses of neurons in cat visual-cortex. *J Physiol (London)*, **440**(Aug.), 659–96.

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

Downing, C. J. 1988. Expectancy and visual-spatial attention: effects on perceptual quality. *J Exp Psychol Hum Percept Perform*, **14**(2), 188–202.

Driver, J., Mcleod, P., & Dienes, Z. 1992. Motion Coherence and Conjunction Search - Implications for Guided Search Theory. *Perception & Psychophysics*, **51**(1), 79–85.

Duncan, J., Ward, R., & Shapiro, K. 1994. Direct measurement of attentional dwell time in human vision. *Nature*, **369**(6478), 313–5.

Engel, S., Zhang, X., & Wandell, B. 1997. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature*, **388**(6637), 68–71.

Eysel, U. T. 1992. Lateral inhibitory interactions in areas 17 and 18 of the cat visual cortex. *Prog Brain Res*, **90**(Bra), 407–22.

Ferster, D. 1986. Orientation selectivity of synaptic potentials in neurons of cat primary visual cortex. *J Neurosci*, **6**(5), 1284–301.

Ferster, D. 1987. Origin of orientation-selective EPSPs in simple cells of cat visual cortex. *J Neurosci*, **7**(6), 1780–91.

Ferster, D. 1988. Spatially opponent excitation and inhibition in simple cells of the cat visual cortex. *J Neurosci*, **8**(4), 1172–80.

Ferster, D., & Jagadeesh, B. 1992. EPSP-IPSP interactions in cat visual cortex studied with in vivo whole-cell patch recording. *J Neurosci*, **12**(4), 1262–74.

Ferster, D., & Koch, C. 1987. Neuronal connections underlying orientation selectivity in cat visual cortex. *Trends Neurosci*, **10**, 187–92.

Fischer, B., Weber, H., Biscaldi, M., Aiple, F., Otto, P., & Stuhr, V. 1993. Separate populations of visually guided saccades in humans: reaction times and amplitudes. *Exp Brain Res*, **92**, 528–541.

Fitzpatrick, D., Lund, J. S., & Blasdel, G. G. 1985. Intrinsic connections of macaque striate cortex: afferent and efferent connections of lamina 4C. *J Neurosci*, **5**(12), 3329–49.

Foley, J. D., van Dam, A., Feiner, S., & Hughes, J. 1990. *Computer Graphics, Principles and Practice (2nd ed.)*. New York, NY: Addison-Wesley.

Foley, J. M. 1994. Human luminance pattern-vision mechanisms: masking experiments require a new model. *J Opt Soc Am A*, **11**(6), 1710–9.

Foley, J. M., & Chen, C. C. 1997. Analysis of the effect of pattern adaptation on pattern pedestal effects: a two-process model. *Vision Res*, **37**(19), 2779–88.

Foley, J. M., & Schwarz, W. 1998. Spatial attention: effect of position uncertainty and number of distractor patterns on the threshold-versus-contrast function for contrast discrimination. *J Opt Soc Am A*, **15**, 1036–1046.

Gallant, J. L., Connor, C. E., & Essen, D. C. Van. 1998. Neural activity in areas V1, V2 and V4 during free viewing of natural scenes compared to controlled viewing. *Neuroreport*, **9**(1), 85–90.

Gandhi, S. P., Heeger, D. J., & Boynton, G. M. 1998. Spatial Attention in Human Primary Visual Cortex. *Investigative Ophtalmology and Visual Science Annual Meeting (ARVO'98)*, **39**(4), 5194.

Garcia-Perez, M. A., & Sierra-Vazquez, V. 1996. Do channels shift their tuning towards lower spatial frequencies in the periphery? *Vision Res*, **36**(20), 3339–72.

Gawne, T. J., Kjaer, T. W., Hertz, J. A., & Richmond, B. J. 1996. Adjacent visual cortical complex cells share about 20% of their stimulus-related information. *Cereb Cortex*, **6**(3), 482–9.

Geisler, W. S., & Albrecht, D. G. 1995. Bayesian analysis of identification performance in monkey visual cortex: nonlinear mechanisms and stimulus certainty. *Vision Res*, **35**(19), 2723–30.

Geisler, W. S., & Albrecht, D. G. 1997. Visual cortex neurons in monkeys and cats: detection, discrimination, and identification. *Vis Neurosci*, **14**(5), 897–919.

Geman, S., & Geman, D. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans Patt Anal Mach Intell*, **6**(6), 721–41.

Gilbert, C. D., & Wiesel, T. N. 1983. Clustered intrinsic connections in cat visual cortex. *J Neurosci*, **3**(5), 1116–33.

Gilbert, C. D., & Wiesel, T. N. 1989. Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *J Neurosci*, **9**(7), 2432–42.

Gilbert, C. D., Das, A., Ito, M., Kapadia, M., & Westheimer, G. 1996. Spatial integration and cortical dynamics. *Proc Natl Acad Sci U S A*, **93**(2), 615–22.

Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. 1998. The representation of visual salience in monkey parietal cortex. *Nature*, **391**(6666), 481–4.

Green, D. M., & Swets, J. A. 1966. *Signal detectability and psychophysics.* New York, NY: Wiley.

Greenlee, M. W. 1992. Spatial frequency discrimination of band-limited periodic targets: effects of stimulus contrast, bandwidth and retinal eccentricity. *Vision Res*, **32**(2), 275–83.

Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit, S., & Anderson, C. H. Jun. 1994. Overcomplete steerable pyramid filters and rotation invariance. *Pages 222–228 of: Proc. IEEE Computer Vision and Pattern Recognition (CVPR), Seattle, WA.*

Hamker, F. H. 1999. The role of feedback connections in task-driven visual search. *In:* von Heinke, D., Humphreys, G. W., & Olson, A. (eds), *Connectionist Models in Cognitive Neuroscience, Proc. of the 5th Neural Computation and Psychology Workshop (NCPW'98).* London: Springer Verlag.

Hata, Y., Tsumoto, T., Sato, H., & Tamura, H. 1991. Horizontal interactions between visual cortical neurones studied by cross-correlation analysis in the cat. *J Physiol (Lond)*, **441**(Sep), 593–614.

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

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

Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. 1996. Computational models of cortical visual processing. *Proc Natl Acad Sci U S A*, **93**(2), 623–7.

Heisenberg, M., & Wolf, R. 1984. *Studies of Brain Function, Vol. 12: Vision in Drosophila.* Berlin, Germany: Springer-Verlag.

Helmholtz, H. 1850/1962. Vol. 3. *In:* Southall, J P C (ed), *Handbuch der Physiologischen Optik.* New York, NY: Dover.

Hikosaka, O., Miyauchi, S., & Shimojo, S. 1996. Orienting a spatial attention–its reflexive, compensatory, and voluntary mechanisms. *Brain Res Cogn Brain Res*, **5**(1-2), 1–9.

Hillstrom, A. P., & Yantis, S. 1994. Visual-Motion and Attentional Capture. *Perception & Psychophysics*, **55**(4), 399–411.

Hirsch, J., & Hylton, R. 1982. Limits of spatial-frequency discrimination as evidence of neural interpolation. *J Opt Soc Am*, **72**(10), 1367–74.

Holt, G. R., & Koch, C. 1997. Shunting inhibition does not have a divisive effect on firing rates. *Neural Comput*, **9**(5), 1001–13.

Horiuchi, T.K., Morris, T.G., Koch, C., & DeWeerth, S.P. 1997. Analog VLSI Circuits for Attention-Based, Visual Tracking. *Pages 706–712 of:* Mozer, M.C., Jordan, M.I., & Petsche, T. (eds), *Neural Information Processing Systems (NIPS*9).* Cambridge, MA: MIT Press.

Horowitz, T. S., & Wolfe, J. M. 1998. Visual search has no memory. *Nature*, **394**(6693), 575–7.

Hubel, D. H., & Wiesel, T. N. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol (London)*, **160**, 106–54.

Itti, L., Braun, J., Lee, D. K., & Koch, C. 1998a. A Model of Early Visual Processing. *Pages 173–9 of:* Jordan, M. I., Kerns, M. J., & Solla, S. A. (eds), *Advances in Neural Information Processing Systems, Vol. 10.* Cambridge, MA: MIT Press.

Itti, L., Koch, C., & Niebur, E. 1998b. A Model of Saliency-Based Visual-Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(11), 1254–9.

Itti, L., Braun, J., Lee, D. K., & Koch, C. 1999. Attentional Modulation of Human Pattern Discrimination Psychophysics Reproduced by a Quantitative Model. *Page (in press) of:* Kearns, M. S., Solla, S. A., & Cohn, D. A. (eds), *Advances in Neural Information Processing Systems, Vol. 11.* Cambridge, MA: MIT Press.

James, W. 1890/1981. *The Principles of Psychology.* Cambridge, MA: Harvard UP.

Kapadia, M. K., Gilbert, C. D., & Westheimer, G. 1994. A quantitative measure for short-term cortical plasticity in human vision. *J Neurosci*, **14**(1), 451–7.

Katz, L. C., Gilbert, C. D., & Wiesel, T. N. 1989. Local circuits and ocular dominance columns in monkey striate cortex. *J Neurosci*, **9**(4), 1389–99.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. 1983. Optimization by Simulated Annealing. *Science*, **220**(4598), 671–80.

Kisvarday, Z. F. 1992. GABAergic networks of basket cells in the visual cortex. *Prog Brain Res*, **90**(Bra), 385–405.

Kisvarday, Z. F., Kim, D. S., Eysel, U. T., & Bonhoeffer, T. 1994. Relationship between lateral inhibitory connections and the topography of the orientation map in cat visual cortex. *Eur J Neurosci*, **6**(10), 1619–32.

Knierim, J. J., & van Essen, D. C. 1992. Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *J Neurophysiol*, **67**(4), 961–80.

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

Koch, C., & Ullman, S. 1985. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, **4**(4), 219–27.

Kustov, A. A., & Robinson, D. L. 1996. Shared neural control of attentional shifts and eye movements. *Nature*, **384**(6604), 74–7.

Kwak, H. W., & Egeth, H. 1992. Consequences of allocating attention to locations and to other attributes. *Percept Psychophys*, **51**(5), 455–64.

Laberge, D., & Buchsbaum, M. S. 1990. Positron Emission Tomographic Measurements of Pulvinar Activity During an Attention Task. *Journal of Neuroscience*, **10**(2), 613–9.

Lamme, V. A. 1995. The neurophysiology of figure-ground segregation in primary visual cortex. *J Neurosci*, **15**(2), 1605–15.

Lee, D. K. 1999. *What You Can See Outside the Focus of Attention (Ph.D. Thesis).* Pasadena, California: California Institute of Technology.

Lee, D. K., Koch, C., & Braun, J. 1997a. Spatial vision thresholds in the near absence of attention. *Vision Res*, **37**(17), 2409–18.

Lee, D. K., Koch, C., & Braun, J. 1997b. Spatial vision thresholds in the near absence of attention. *Vision Res*, **37**(17), 2409–18.

Lee, D. K., Itti, L., Koch, C., & Braun, J. 1999a. Attention activates winner-take-all competition among visual filters. *Nat Neurosci*, **2**(4), 375–81.

Lee, D. K., Itti, L., Koch, C., & Braun, J. 1999b. Attention activates winner-take-all competition among visual filters. *Nat Neurosci*, **2**(4), 375–81.

Legge, G. E., & Foley, J. M. 1980. Contrast masking in human vision. *J Opt Soc Am*, **70**(12), 1458–71.

Leventhal, A.G. 1991. *The Neural Basis of Visual Function (Vision and Visual Dysfunction Vol. 4).* Boca Raton, FL: CRC Press.

Levitt, J. B., & Lund, J. S. 1997. Contrast dependence of contextual effects in primate visual cortex. *Nature*, **387**(6628), 73–6.

Lu, Z. L., & Dosher, B. A. 1998. External noise distinguishes attention mechanisms. *Vision Res.*, **38**, 1183–1198.

Luck, S. J., Chelazzi, L., Hillyard, S. A., & Desimone, R. 1997. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J Neurophysiol*, **77**(1), 24–42.

Luschow, A., & Nothdurft, H. C. 1993. Pop-out of Orientation but no pop-out of Motion at Isoluminance. *Vision Research*, **33**(1), 91–104.

Magnussen, S., Greenlee, M. W., & Thomas, J. P. 1996. Parallel processing in visual short-term memory. *J Exp Psychol Hum Percept Perform*, **22**(1), 202–12.

Malach, R. 1994. Cortical columns as devices for maximizing neuronal diversity. *Trends Neurosci*, **17**(3), 101–4.

Malach, R., Amir, Y., Harel, M., & Grinvald, A. 1993. Relationship between intrinsic connections and functional architecture revealed by optical imaging and in vivo targeted biocytin injections in primate striate cortex. *Proc Natl Acad Sci U S A*, **90**(22), 10469–73.

Malik, J., & Perona, P. 1990. Preattentive texture discrimination with early vision mechanisms. *J Opt Soc Am [A]*, **7**(5), 923–32.

Mato, G., & Sompolinsky, H. 1996. Neural network models of perceptual learning of angle discrimination. *Neural Comput*, **8**(2), 270–99.

Maunsell, J. H. 1995. The brain's visual world: representation of visual targets in cerebral cortex. *Science*, **270**(5237), 764–9.

McAdams, C. J., & Maunsell, J. H. R. 1996a. Attention enhances neuronal reponses without altering orientation selectivity in macaque area V4. *In: Neuroscience abstract.*

McAdams, C. J., & Maunsell, J. H. R. 1996b. Attention enhances neuronal responses without altering orientation selectivity in macaque area V4. *Page 475.2 of: Soc Neurosci Abstr*, vol. 22.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., & Teller, E. 1953. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Milanese, R., Gil, S., & Pun, T. 1995. Attentive Mechanisms for Dynamic and Static Scene Analysis. *Opt Eng*, **34**(8), 2428–34.

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

Motter, B. C. 1993. Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *J Neurophysiol*, **70**(3), 909–19.

Motter, B. C. 1994a. Neural correlates of attentive selection for color or luminance in extrastriate area V4. *J Neurosci*, **14**(4), 2178–89.

Motter, B. C. 1994b. Neural correlates of feature selective memory and pop-out in extrastriate area V4. *J Neurosci*, **14**(4), 2190–9.

Motter, B. C., & Belky, E. J. 1998. The guidance of eye movements during active visual search. *Vision Res*, **38**(12), 1805–15.

Müller, H J, & Findlay, J. 1987a. Sensitivity and criterion effects in the spatial cuing of visual attention. *Perception & Psychophysics*, **42**, 383–399.

Müller, H. J., & Findlay, J. M. 1987b. Sensitivity and criterion effects in the spatial cuing of visual attention. *Percept Psychophys*, **42**(4), 383–99.

Müller, H. J., & Humphreys, G. W. 1991. Luminance-increment detection: capacity-limited or not? *J Exp Psychol Hum Percept Perform*, **17**(1), 107–24.

Nachmias, J., & Sansbury, R. V. 1974. Letter: Grating contrast: discrimination may be better than detection. *Vision Res*, **14**(10), 1039–42.

Nakayama, K., & Mackeben, M. 1989. Sustained and transient components of focal visual attention. *Vision Res*, **29**(11), 1631–47.

Nelson, S. B. 1991. Temporal interactions in the cat visual system. I. Orientation-selective suppression in the visual cortex. *J Neurosci*, **11**(2), 344–56.

Niebur, E., & Koch, C. 1994. A model for the neuronal implementation of selective visual attention based on temporal correlation among neurons. *J Comput Neurosci*, **1**(1-2), 141–58.

Niebur, E., & Koch, C. 1996. Control of Selective Visual Attention: Modeling the 'Where' Pathway. *Pages 802–808 of:* Touretzky, D.S., Mozer, M.C., & Hasselmo, M.E. (eds), *Neural Information Processing Systems (NIPS*8).* Cambridge, MA: MIT Press.

Niebur, E., & Koch, C. 1998. Computational architectures for attention. *Pages 163–186 of:* Parasuraman, R. (ed), *The attentive brain.* Cambridge, MA: MIT Press.

Niyogi, P., Girosi, F., & Poggio, T. 1998. Incorporating Prior Information in Machine Learning by Creating Virtual Examples. *Proceedings of the IEEE*, **86**(11), 2196–209.

Noton, D., & Stark, L. 1971. Scanpaths in eye movements during pattern perception. *Science*, **171**(968), 308–11.

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

Olzak, L. A., & Thomas, J. P. 1991. When orthogonal orientations are not processed independently. *Vision Res*, **31**(1), 51–7.

O'Regan, J. K., Rensink, R. A., & Clark, J. J. 1999. Change-blindness as a result of 'mudsplashes'. *Nature*, **398**(6722), 34.

Palmer, J. 1995. Attention in visual search: distinguishing four causes of set-size effects. *Curr Dir Psychol Sci*, **4**, 118–123.

Pashler, H. 1997. *The Psychology of Attention.* Cambridge, MA: MIT Press.

Phillips, G. C., & Wilson, H. R. 1984. Orientation bandwidths of spatial mechanisms measured by masking. *J Opt Soc Am [A]*, **1**(2), 226–32.

Poggio, T. 1997. Image Representations for Visual Learning. *Lecture Notes in Computer Science*, **1206**, 143.

Polat, U., & Sagi, D. 1994a. The architecture of perceptual spatial interactions. *Vision Res*, **34**(1), 73–8.

Polat, U., & Sagi, D. 1994b. Spatial interactions in human vision: from near to far via experience-dependent cascades of connections. *Proc Natl Acad Sci U S A*, **91**(4), 1206–9.

Pollen, D. A., & Ronner, S. F. 1981. Phase relationships between adjacent simple cells in the visual cortex. *Science*, **212**(4501), 1409–11.

Posner, M. I., Cohen, Y., & Rafal, R. D. 1982. Neural systems control of spatial orienting. *Philos Trans R Soc Lond B Biol Sci*, **298**(1089), 187–98.

Pouget, A., Zhang, K., Deneve, S., & Latham, P. E. 1998. Statistically Efficient Estimation using Population Coding. *Neural Comput*, **10**, 373–401.

Press, W. A., Knierim, J. J., & Van Essen, D. C. 1994. Neuronal correlates of attention to texture patterns in macaque striate cortex. *Soc Neurosci Abstr*, **20**, 838.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992. *Numerical Recipes in C.* Cambridge, MA: Cambridge University Press.

Quick, R. F. 1974. A vector-magnitude model of contrast detection. *Kybernetik*, **16**, 1299–302.

Ramoa, A. S., Shadlen, M., Skottun, B. C., & Freeman, R. D. 1986. A comparison of inhibition in orientation and spatial frequency selectivity of cat visual cortex. *Nature*, **321**(6067), 237–9.

Rao, R. P. N., & Ballard, D. H. 1995. An Active Vision Architecture Based on Iconic Representations. *Artificial Intelligence*, **78**(1-2), 461–505.

Reinagel, P., & Zador, A. M. 1999. Natural scene statistics at the centre of gaze. *Network: Comput. Neural Syst.*, **10**, 341–350.

Reynolds, J., Pasternak, T., & Desimone, R. 1994a. Attention Increases Contrast Sensitivity of cells in Macaque Area V4. *Soc Neurosci Abstr*, **22**, 475.3.

Reynolds, J., Chelazzi, L., Luck, S., & Desimone, R. 1994b. Sensory interactions and effects of selective spatial attention in macaque area V2. *Soc Neurosci Abstr*, **20**, 1054.

Reynolds, J., Nicholas, J., Chelazzi, L, & Desimone, R. 1995. Spatial attention protects macaque V2 and V4 cells from the influence of non-attended stimuli. *Soc Neurosci Abstr*, **21**, 1759.

Reynolds, J., Pasternak, T., & Desimone, R. 1997. Attention Increases Contrast Sensitivity of Cells in Macaque Area V4. *Inv Ophth Vis Sci (Suppl)*, **38**, 3206.

Robinson, D. L., & Petersen, S. E. 1992. The pulvinar and visual salience. *Trends Neurosci*, **15**(4), 127–32.

Rockland, K. S., & Lund, J. S. 1983. Intrinsic laminar lattice connections in primate visual cortex. *J Comp Neurol*, **216**(3), 303–18.

Rockland, K. S., Andresen, J., Cowie, R. J., & Robinson, D. L. 1999. Single axon analysis of pulvinocortical connections to several visual areas in the macaque. *J Comp Neurol*, **406**(2), 221–50.

Roelfsema, P. R., Lamme, V. A., & Spekreijse, H. 1998. Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, **395**, 376–381.

Saarinen, J., & Julesz, B. 1991. The speed of attentional shifts in the visual field. *Proc Natl Acad Sci U S A*, **88**(5), 1812–4.

Sachs, M. B., Nachmias, J., & Robson, J. G. 1971. Spatial-frequency channels in human vision. *J Opt Soc Am*, **61**(9), 1176–86.

Salzman, C. D., & Newsome, W. T. 1994. Neural mechanisms for forming a perceptual decision. *Science*, **264**(5156), 231–7.

Scharf, L. L. 1991. *Statistical Signal Processing: Detection, Estimation and Time-Series Analysis*. Reading, MA: Addison-Wesley.

Seung, H. S., & Sompolinsky, H. 1993. Simple models for reading neuronal population codes. *Proc Natl Acad Sci U S A*, **90**(22), 10749–53.

Shadlen, M. N., Britten, K. H., Newsome, W. T., & Movshon, J. A. 1996. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *J Neurosci*, **16**(4), 1486–510.

Sheliga, B. M., Riggio, L., & Rizzolatti, G. 1994. Orienting of attention and eye movements. *Exp Brain Res*, **98**(3), 507–22.

Shepherd, M., Findlay, J. M., & Hockey, R. J. 1986. The relationship between eye movements and spatial attention. *Q J Exp Psychol*, **38**, 475–491.

Sillito, A. M. 1979. Inhibitory mechanisms influencing complex cell orientation selectivity and their modification at high resting discharge levels. *J Physiol (Lond)*, **289**(Apr), 33–53.

Sillito, A. M., & Jones, H. E. 1996. Context-dependent interactions and visual processing in V1. *J Physiol Paris*, **90**(3-4), 205–9.

Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J., & Davis, J. 1995. Visual cortical mechanisms detecting focal orientation discontinuities. *Nature*, **378**(6556), 492–6.

Simoncelli, E. P., Freeman, W. T., Adelson, E. H., & Heeger, D. J. 1992. Shiftable Multiscale Transforms. *Ieee Transactions on Information Theory*, **38**(2), 587–607.

Simons, D. J., & Levin, D. T. 1997. Failure to Detect Changes to Attended Objects. *Investigative Ophthalmology & Visual Science*, **38**(4), 3273.

Skottun, B. C., Bradley, A., Sclar, G., Ohzawa, I., & Freeman, R. D. 1987. The effects of contrast on visual orientation and spatial frequency discrimination: a comparison of single cells and behavior. *J Neurophysiol*, **57**(3), 773–86.

Snippe, H. P., & Koenderink, J. J. 1992. Information in channel-coded systems: correlated receivers. *Biol Cybern*, **67**(2), 183–90.

Snowden, R. J., & Hess, R. F. 1992. Temporal frequency filters in the human peripheral visual field. *Vision Res*, **32**(1), 61–72.

Softky, W. R., & Koch, C. 1993. The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *J Neurosci*, **13**(1), 334–50.

Solomon, J. A., Lavie, N., & Morgan, M. J. 1997. Contrast discrimination function: spatial cuing effects. *J Opt Soc Am A*, **14**, 2443–2448.

Somers, D. C., Nelson, S. B., & Sur, M. 1995. An emergent model of orientation selectivity in cat visual cortical simple cells. *J Neurosci*, **15**(8), 5448–65.

Sompolinsky, H., & Shapley, R. 1997. New perspectives on the mechanisms for orientation selectivity. *Curr Opin Neurobiol*, **7**(4), 514–22.

Spitzer, H., Desimone, R., & Moran, J. 1988. Increased attention enhances both behavioral and neuronal performance. *Science*, **240**(4850), 338–40.

Suarez, H., Koch, C., & Douglas, R. 1995. Modeling direction selectivity of simple cells in striate visual cortex within the framework of the canonical microcircuit. *J Neurosci*, **15**(10), 6700–19.

Teich, M. C., Turcott, R. G., & Siegel, R. M. 1996. Temporal Correlation in Cat Striate-Cortex Neural Spike Trains. *IEEE Eng Med Biol*, **Sept-Oct**, 79–87.

Thibos, L. N., Still, D. L., & Bradley, A. 1996. Characterization of spatial aliasing and contrast sensitivity in peripheral vision. *Vision Res*, **36**(2), 249–58.

Thomas, J. P., & Olzak, L. A. 1997. Contrast gain control and fine spatial discriminations. *J Opt Soc Am A*, **14**(9), 2392–405.

Toepfer, C., Wende, M., Baratoff, G., & Neumann, H. 1998. Robot Navigation by Combining Central and Peripheral Optical Flow Detection on a Space-Variant Map. *Pages 1804–7 of: Proc. 14th Int. Conf. on Pattern Recognition (ICPR'98), Brisbane, Australia.*

Toet, A., Bijl, P., Kooi, F. L., & Valeton, J. M. 1998. *A high-resolution image dataset for testing search and detection models (TNO-TM-98-A020).* Soesterberg, The Netherlands: TNO Human Factors Research Institute.

Tootell, R. B., Hamilton, S. L., Silverman, M. S., & Switkes, E. 1988. Functional anatomy of macaque striate cortex. I. Ocular dominance, binocular interactions, and baseline conditions. *J Neurosci*, **8**(5), 1500–30.

Toyama, K., Kimura, M., & Tanaka, K. 1981. Cross-Correlation Analysis of Interneuronal Connectivity in cat visual cortex. *J Neurophysiol*, **46**(2), 191–201.

Treisman, A. 1988. Features and objects: the fourteenth Bartlett memorial lecture. *Q J Exp Psychol [A]*, **40**(2), 201–37.

Treisman, A., & Gormican, S. 1988. Feature analysis in early vision: evidence from search asymmetries. *Psychol Rev*, **95**(1), 15–48.

Treisman, A. M., & Gelade, G. 1980. A feature-integration theory of attention. *Cognit Psychol*, **12**(1), 97–136.

Treue, S., & Maunsell, J. H. 1996. Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature*, **382**(6591), 539–41.

Treue, S., & Trujillo, J. C. Martinez. 1999. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, **399**(6736), 575–9.

Ts'o, D. Y., Gilbert, C. D., & Wiesel, T. N. 1986. Relationships between horizontal interactions and functional architecture in cat striate cortex as revealed by cross-correlation analysis. *J Neurosci*, **6**(4), 1160–70.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y. H., Davis, N., & Nuflo, F. 1995. Modeling Visual-Attention via Selective Tuning. *Artificial Intelligence*, **78**(1-2), 507–45.

Tyler, C. W. 1997. Colour bit-stealing to enhance the luminance resolution of digital displays on a single pixel basis. *Spat Vis*, **10**(4), 369–77.

Verghese, P., & Stone, L. S. 1995. Combining speed information across space. *Vision Res*, **35**(20), 2811–23.

Vogels, R., & Orban, G. A. 1990. How well do response changes of striate neurons signal differences in orientation: a study in the discriminating monkey. *J Neurosci*, **10**(11), 3543–58.

Wagenaar, W. A. 1969. Note on the construction of digram-balanced Latin squares. *Psychol Bull*, **72**, 384–86.

Watson, A. B. 1990. Neural Contrast Sensitivity. *Pages 95–108 of:* Landy, M. S., & Movshon, J. A. (eds), *Computational Models of Visual Processing*. Cambridge, MA: MIT Press.

Watson, A. B., & Pelli, D. G. 1983. QUEST: a Bayesian adaptive psychometric method. *Percept Psychophys*, **33**(2), 113–20.

Watt, R. J., & Morgan, M. J. 1985. A theory of the primitive spatial code in human vision. *Vision Res*, **25**, 1661–1674.

Weibull, W. A. 1951. A statistical distribution function of wide applicability. *J Appl Mechan*, **18**, 292–297.

Weliky, M., Kandler, K., Fitzpatrick, D., & Katz, L. C. 1995. Patterns of excitation and inhibition evoked by horizontal connections in visual cortex share a common relationship to orientation columns. *Neuron*, **15**(3), 541–52.

Wilkinson, F., Wilson, H. R., & Ellemberg, D. 1997. Lateral interactions in peripherally viewed texture arrays. *J Opt Soc Am A*, **14**(9), 2057–68.

Wilson, H. R. 1980a. A transducer function for threshold and suprathreshold human vision. *Biol Cybern*, **38**, 171–178.

Wilson, H. R. 1980b. A transducer function for threshold and suprathreshold human vision. *Biol Cybern*, **38**(3), 171–8.

Wilson, H. R. 1986. Responses of spatial mechanisms can explain hyperacuity. *Vision Res*, **26**(3), 453–69.

Wilson, H. R. 1991. Model of peripheral and amblyopic hyperacuity. *Vision Res*, **31**(6), 967–82.

Wilson, H. R. 1993. Nonlinear processes in visual pattern discrimination. *Proc Natl Acad Sci U S A*, **90**(21), 9785–90.

Wilson, H. R., & Bergen, J. R. 1979. A four mechanism model for threshold spatial vision. *Vision Res*, **19**(1), 19–32.

Wilson, H. R., & Gelb, D. J. 1984. Modified line-element theory for spatial-frequency and width discrimination. *J Opt Soc Am [A]*, **1**(1), 124–31.

Wilson, H. R., & Humanski, R. 1993. Spatial frequency adaptation and contrast gain control. *Vision Res*, **33**(8), 1133–49.

Wilson, H. R., & Wilkinson, F. 1997. Evolving concepts of spatial channels in vision: from independence to nonlinear interactions. *Perception*, **26**, 939–960.

Wilson, H. R., McFarlane, D. K., & Phillips, G. C. 1983. Spatial frequency tuning of orientation selective units estimated by oblique masking. *Vision Res*, **23**(9), 873–82.

Wilson, H. R., Levi, D., Maffei, L., Rovamo, J., & DeValois, R. 1990. The Pereception of Form: Retina to Striate Cortex. *Pages 231–272 of:* Spillmann, L., & Werner, J. S. (eds), *Visual Perception: The Neurophysiological Foundations*. San Diego, CA: Academic Press.

Wolfe, J. M. 1994. Visual search in continuous, naturalistic stimuli. *Vision Res*, **34**(9), 1187–95.

Yarbus, A. L. 1967. *Eye Movements and Vision*. New York: Plenum Press.

Yuille, A. L., & Grzywacz, N. M. 1989. A Mathematical-Analysis of the Motion Coherence Theory. *International Journal of Computer Vision*, **3**(2), 155–75.

Zenger, B., & Sagi, D. 1996. Isolating excitatory and inhibitory nonlinear spatial interactions involved in contrast detection. *Vision Res*, **36**(16), 2497–513.

Zipser, K., Lamme, V. A., & Schiller, P. H. 1996. Contextual modulation in primary visual cortex. *J Neurosci*, **16**(22), 7376–89.

Zohary, E., Shadlen, M. N., & Newsome, W. T. 1994. Correlated neuronal discharge rate and its implications for psychophysical performance [published erratum appears in Nature 1994 Sep 22;371(6495):358]. *Nature*, **370**(6485), 140–3.