Grouping mechanisms for object-based vision and attention

by

Brian H. Hu

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

August, 2017

Abstract

The visual brain faces the difficult task of reconstructing a three-dimensional (3D) world from two-dimensional (2D) images projected onto the two retinae. In doing so, visual information is organized in terms of objects in 3D space, and this organization is the basis for visual perception. In complex visual scenes, both the foreground and the background are rich in features of different types. The brain must find a way to group together the features that belong to objects on the foreground and distinguish them from features in the background.

The goal of this thesis is to understand how the neural circuits in primate cortex accomplish this task using grouping mechanisms for object-based vision and attention. Through computational modeling, I show that grouping mechanisms are fundamental for linking early feature representations to tentative perceptual objects known as proto-objects. Previous models on the neural coding of border ownership have identified a plausible network architecture for proto-object based perceptual organization. I extend these models to explain how the same grouping model framework can be used to perform contour integration, border-ownership assignment, grouping

ABSTRACT

of 3D surfaces, and computation of 3D visual saliency. My models offer several falsifiable predictions which can be tested in future experiments. My models also clarify how top-down attention interfaces with the neural circuits responsible for grouping together the features of an object. Overall, the models developed address the important question of how visual features are grouped into 2D and 3D object representations.

Primary Reader: Ernst Niebur

Secondary Reader: Rüdiger von der Heydt

Acknowledgments

I had the privilege of sailing with Ernst to the neuroscience retreat one year. As I was sitting on the boat, I thought about how the trip was really a metaphor for my PhD. I have had opportunities to steer the boat (sometimes in the wrong direction!) and I have gone through both good and bad weather. Through it all, I am thankful to have had Ernst as my captain.

I would like to thank my family for their continued love and support. I thank my wife, Mingming, and my daughter, Katherine, for bearing with me in these final months. I thank my parents, Benjamin and Jennifer, for the sacrifices they made for me and my siblings. I thank my brother Blair for his moral support and my sister Joy for being an inspiration to me.

I would also like to thank my thesis committee members, Rüdiger and Kechen, for their help and guidance during my PhD. Finally, I would like to thank my friend Matthew, for all the lunches and discussions we shared together, and my labmates Danny and Grant, who made each day in lab an interesting one.

Soli Deo Gloria

Dedication

This thesis is dedicated to Joy, who first got me interested in the study of vision. Your perseverance in the face of adversity has been my inspiration.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Segmentation and figure-ground organization

The task of partitioning an image into regions bounded by contours (segmentation) and the task of assigning border ownership of these contours to either the foreground or the background (figure-ground organization) are important first steps in achieving image understanding. Both tasks require the brain to keep track of which regions and contours belong to which objects. This is known as the binding problem, as it is not clear how the features of an object are bound together (Treisman, 1996). One class of models relies on the fast temporal coding structures of spike trains (Singer, 1999), but experimental evidence is controversial (Thiele and Stoner, 2003; Roelfsema et al.,

2004; Dong et al., 2008). Another solution involves differential neural activity, where the neurons responding to the features of an object show increased firing compared with neurons responding to the background. This response enhancement is known as figure-ground modulation (FGM), and was first observed in primary visual cortex (V1) for texture-defined figures (Lamme, 1995). Similar results have been found using other tasks and techniques, including more recent voltage-sensitive dye imaging of populations of neurons during a contour grouping task (Gilad et al., 2013).

However, this solution only works if there is a single object in the foreground, as multiple objects each labeled with higher neural activity could be interpreted as parts of a single object. Furthermore, each neuron's firing rate is inherently ambiguous, as higher activity could be due to labeling with FGM or because the neuron's preferred feature falls within its receptive field (RF). As a result, the binding problem cannot be solved with models that only represent object information in terms of enhanced neural activity in early visual areas (Niebur, 2000).

1.2 The role of cortical feedback

Early computational models (Stemmler et al., 1995b) and experimental studies (Simonotto et al., 1997; Polat et al., 1998; Chatterjee et al., 2011; Xie et al., 2014) put forth the view that horizontal connections are essentially static, or varying over time scales given by ontogenetic development or neuronal plasticity. Such structures could

thus implement overall statistics of natural scenes (like circular structures, e.g. Sigman et al., 2001) but they could not flexibly represent the myriad of instantaneously present and constantly changing visual shapes observed during perception of dynamic scenes. This view has been considerably enlarged over the last decade or so, and what is emerging is a view in which the lateral connections are in place but can be actively and rapidly modulated by top-down connectivity from higher areas.

The degree of collinear facilitation observed in V1 is strongly context-dependent, and can change with the behavioral task (Li et al., 2004, 2006) as well as perceptual learning (Li et al., 2008; Yan et al., 2014). As a result, feedback connections from higher areas may play an important role in shaping the responses of neurons in early visual areas. In fact, simultaneous neural recordings from areas V1 and V4 during two different figure-ground segregation tasks show that V4 is intimately involved in the FGM process (Poort et al., 2012; Chen et al., 2014). In these studies, the FGM signal appears first in V4 and is then fed back to V1, with a delay representing recurrent processing. Additional studies of curve-tracing (Roelfsema et al., 1998) and border-ownership (Zhou et al., 2000; Qiu et al., 2007; Zhang and von der Heydt, 2010) further demonstrate that feedback mechanisms are necessary for explaining FGM in the presence of multiple objects. However, essential questions still remain about the nature of the interactions between and within different cortical areas. One of the goals of this thesis is to understand how early-level feature information about an object is combined with global context information about the object in a synergistic way in

order to generate FGM.

1.3 The role of attention

Behavioral studies have shown that attention can be directed to objects (Egly et al., 1994) and electrophysiological results demonstrate that attention can act as a top-down signal which influences FGM (Qiu et al., 2007; Poort et al., 2012). In an ambiguous figure-ground display, attending to one region increases the probability that this region is perceived as figure (Driver and Baylis, 1996; Vecera et al., 2004). Spatial attention, which has been extensively studied, acts like a "spotlight" that enhances neural responses within the focus of attention and suppresses responses outside (Motter, 1993). Attention can also operate in a feature-based or object-based manner. Feature-based attention acts broadly across the visual scene and increases the responses of all components that share similar feature attributes (e.g. color, orientation, or direction of movement) with the attended component (Treue, 1999). Object-based attention highlights all the parts of an object, also encompassing all the features that belong to the object (Roelfsema et al., 1998; Schoenfeld et al., 2014). Attention has been found to modulate border ownership in an object-based manner (Qiu et al., 2007).

Border ownership is a property of many neurons in V2 which encodes the side to which an object border belongs relative to their RF (Zhou et al., 2000). To explain

these results, Craft et al. (2007) proposed a model in which populations of grouping neurons explicitly represent (in their firing rates) the perceptual organization of the visual scene. Grouping neurons are reciprocally connected to border-ownership selective (BOS) neurons through feedforward and feedback connections. Attention broadly targets grouping neurons, which can then modulate the activity of BOS neurons through feedback (Mihalas et al., 2011). A feedforward version of this model has been applied to natural images, where it outperforms other models in predicting the location of eye fixations (Russell et al., 2014).

1.4 Grouping of 3D surfaces

Grouping mechanisms are important not only for piecing together object contours, but also for providing a structure for selectively attending to groups of objects (Treisman and Gelade, 1980). Supported by extensive psychophysical data, Nakayama, He, and Shimojo (1995) proposed that surface representations play a key role in intermediate-level vision. For example, by selectively attending to a surface in 3D space, subjects can perform efficient search for a conjunction target (Nakayama and Silverman, 1986). In a separate cueing experiment, attention was shown to spread automatically across surfaces (He and Nakayama, 1995). These abilities indicate powerful mechanisms for grouping objects into surfaces in 3D space, and suggest that structuring the world in terms of surfaces might be an ecologically important

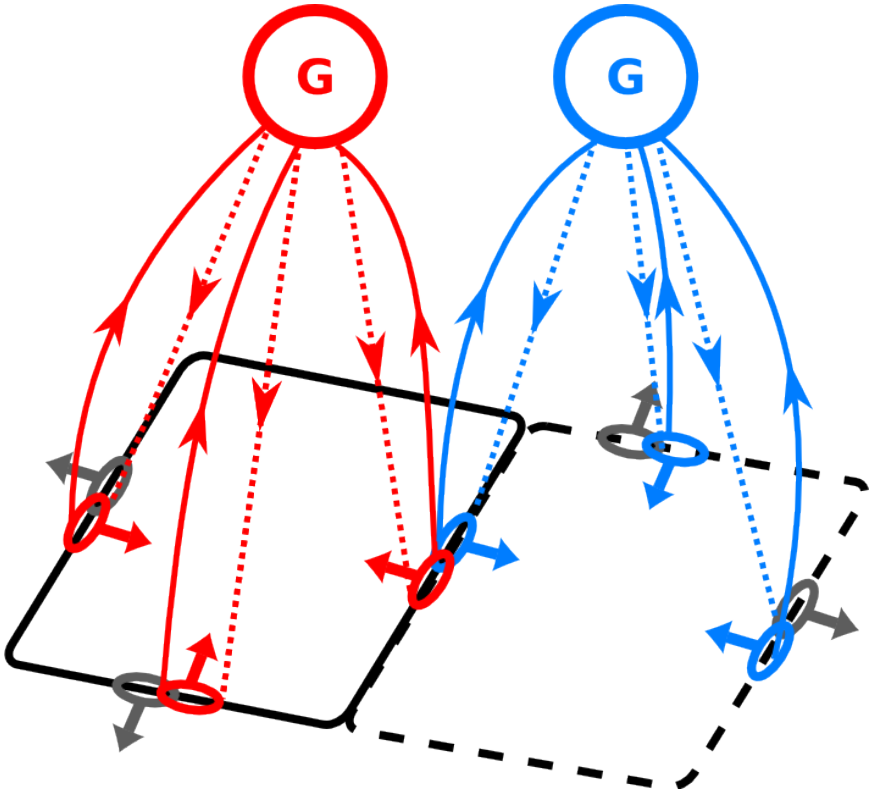


Figure 1.1: Overview of the grouping model

function (*e.g.* for locomotion along the ground plane, reaching for objects along a table top *etc.*). Modeling the grouping of 3D surfaces will provide insight into the neural circuits that represent surfaces, filling a critical gap in our understanding of intermediate-level vision.

1.5 Overview of the grouping model

Several models have been proposed (Zhaoping, 2005; Sakai and Nishimura, 2006; Craft et al., 2007; Layton et al., 2012) to describe how a neuron's border ownership selectivity can be modulated by visual input far away from its classical RF. In the

grouping model (Craft et al., 2007), BOS neurons participate in neural circuits that define perceptual objects early-on in processing (Figure 1.1). An object border activates a pair of orientation-selective BOS neurons whose RFs are shown as ellipses. The arrows on the RFs point towards the preferred direction of the neuron, indicating where the object is relative to the neuron's RF. BOS neurons activated by the solid-line square (red ellipses) excite the appropriate grouping neuron (red G in circle) through feedforward connections (red solid lines). The grouping neuron in turn enhances the activity of the same BOS neurons through feedback connections (red dotted lines). This type of facilitatory feedback may be mediated by NMDAR channels, which allow gating of sensory input by top-down signals (Palmer et al., 2014; Wagatsuma et al., 2016). Neurons consistent with other objects (e.g. the dashed-line square) project to other grouping neurons (blue G in circle in this case). Presence of the solid-line square increases the firing rate of the red grouping neuron over that of the blue (and other) grouping neurons since the latter receive less feedforward input. As a consequence, the red grouping neuron provides more feedback to the red BOS neurons than the blue and gray BOS neurons would receive from their respective grouping neurons. Likewise, presence of the dashed-line square increases the firing rates of all blue neurons over those of gray and red neurons. Thus, an object border is represented by two BOS neurons whose relative activity codes for the side of ownership. The relative difference in firing rate is also known as the BOS signal (Zhou et al., 2000).

The BOS signal appears ˜25 ms after the visual response to an oriented edge, and the delay is essentially independent of object size (Zhou et al., 2000). This constant delay is consistent with a model in which grouping neurons of different sizes integrate local edge signals, and by feedback enhance the same edge signals (Craft et al., 2007). Attention enhances a BOS neuron's response when an object is on the neuron's preferred side, but has a suppressive effect if the object is on the non-preferred side (Qiu et al., 2007). This asymmetry is consistent with a model in which top-down attention targets grouping neurons, which then modulate the activity of BOS neurons through feedback (Mihalas et al., 2011). Additional support for the grouping model comes from observations of short-term memory of BOS signals (O'Herron and von der Heydt, 2009) and remapping of BOS signals across saccades and object movements (O'Herron and von der Heydt, 2013). These findings are difficult to explain with models that only represent object information in terms of neural activity in early visual areas. This strongly points to neural circuits that employ populations of neurons which explicitly represent (*i.e.* in their firing rate) the organization of the visual scene.

1.6 Summary of thesis

The work presented in this thesis deepens and extends our understanding of the neural mechanisms of FGM. In Chapter 1, we introduce background information

needed to understand the physiology and previous modeling experiments. In Chapter 2, we propose a quantitative neural model of grouping constrained by physiological data. We validate the model by reproducing several experimental results related to contour integration and border-ownership assignment. This work is published (Hu and Niebur, 2017). In Chapter 3, we extend this model to natural scenes, and our model results are quantitatively compared with both neurophysiological results and human-annotated figure-ground labels (Berkeley Segmentation Dataset). Beginning with Chapter 4, we shift our focus to the representation of 3D information in the visual system. First, we show that a grouping model can reproduce results from a set of psychophysical experiments where attention had to be directed to surfaces. This work is published (Hu et al., 2015). We then show that 3D surfaces can be represented by a feedforward, linear combination of basis functions whose response properties are similar to those of disparity-selective neurons commonly found in early visual cortex. In Chapter 5, we propose a model of 3D visual saliency and show that the added depth information improves saliency prediction. This work is published (Hu et al., 2016).

Chapter 2

Contour integration and border-ownership assignment

2.1 Introduction

Gestalt psychologists recognized the importance of the whole in influencing perception of the parts when they laid out several principles ("Gestalt laws") for perceptual organization (Wertheimer, 1923; Koffka, 1935). Contour integration, the linking of line segments into contours, and figure-ground segregation, the segmenting of objects from background, are fundamental components of this process. Both require combining local, low-level and global, high-level information that is represented in different areas of the brain in order to segment the visual scene. The interaction between feedforward and feedback streams carrying this information, as well as the contribution

of top-down influences such as attentional selection, are not well understood.

The contour integration process begins in primary visual cortex (V1), where the responses of orientation selective neurons can be modulated by placing collinear stimuli outside the receptive fields (RFs) of these neurons (Stemmler et al., 1995a; Polat et al., 1998). Contextual interactions between V1 neurons have often been summarized using a "local association field," where collinear contour elements excite each other and noncollinear elements inhibit each other (Ullman et al., 1992; Field et al., 1993). Results from neuroanatomy lend support to these ideas, as the lateral connections within V1 predominantly link similar-orientation cortical columns (Gilbert and Wiesel, 1989; Bosking et al., 1997; Stettler et al., 2002). Computational models based on these types of local interactions have successfully explained the ability of V1 neurons to extract contours from complex backgrounds (Li, 1998; Yen and Finkel, 1998; Piëch et al., 2013). While some of these models also incorporate feedback connections, the mechanisms by which higher visual areas construct the appropriate feedback signals and target early feature neurons are not clearly specified. One of the main purposes of this chapter is to introduce concrete neural circuitry that makes this recurrent structure explicit, thereby allowing us to make quantitative predictions and compare model predictions with experimental data.

Segmenting an image into regions corresponding to objects requires not only finding the contours in the image but also determining which contours belong to which objects. Border-ownership selective cells that have been found in early visual ar-

eas, predominantly in secondary visual cortex (V2), appear to be dedicated to this task (Zhou et al., 2000). Border-ownership selective cells encode where an object is located relative to their RFs. When the edge of a figure is presented in its RF, a border-ownership cell will respond with different firing rates depending on where the figure is located. For example, a vertical edge can belong either to an object on the left or on the right of the RF. A border-ownership selective cell will respond differently to these two cases, firing at a higher rate when the figure is located on its "preferred" side, even though the stimulus within its RF may be identical. For a more detailed operational definition of how border-ownership selectivity is determined experimentally, see Section 2.2.4. Border-ownership coding has been studied using a wide variety of artificial stimuli, including those defined by luminance contrast, color contrast, figure outlines (Zhou et al., 2000), motion (von der Heydt et al., 2003), disparity (Zhou et al., 2000; Qiu and von der Heydt, 2005), and transparency (Qiu and von der Heydt, 2007) as well as, more recently, with faces (Hesse and Tsao, 2016) and within complex natural scenes (Williford and von der Heydt, 2014, 2016).

To explain this phenomenon, some computational models assume that image context integration is achieved by propagation of neural activity along horizontal connections within early visual areas. Border-ownership information could be generated from the asymmetric organization of surrounds (Nishimura and Sakai, 2004, 2005; Sakai et al., 2012) or from a diffusion-like process within the image representation (Grossberg, 1994, 1997; Baek and Sajda, 2005; Kikuchi and Akashi, 2001; Pao et al., 1999;

Zhaoping, 2005). However, these models have difficulties explaining the fast establishment of border ownership which appears about 25ms after the first stimulus response (Zhou et al., 2000). Propagation along horizontal fibers over the distances used in the experiments would imply a delay of at least $\approx$ 70ms (Girard et al., 2001, based on the conduction velocity of horizontal fibers in primate V1 cortex; we are not aware of corresponding data for V2). Furthermore, such models are difficult to reconcile with the observation that the time course of border-ownership coding is largely independent of figure size (Sugihara et al., 2011).

An alternative computational model involves populations of grouping (G) cells which explicitly represent (in their firing rates) the perceptual organization of the visual scene (Schütze et al., 2003; Craft et al., 2007). These cells are reciprocally connected to border-ownership selective (B) neurons through feedforward and feedback connections. The combination of grouping cells and the cells signaling local features represents the presence of a "proto-object"(Rensink, 2000), resulting in a structured perceptual organization of the scene. Among the operations that can be performed efficiently in the organized scene are tasks that require attention to objects. In our model, attention to an object targets the grouping neurons representing it, rather than, *e.g.* , all low-level features within a visual area that is defined purely spatially (like everything within a certain distance from the center of attention). Therefore attention is directed to proto-objects, resulting in the modulation of B cell activity through feedback from grouping cells (Mihalas et al., 2011). This proto-object based

approach is consistent with psychophysical and neurophysiological studies (*e.g.* Duncan, 1984; Egly et al., 1994; Scholl, 2001; Kimchi et al., 2007; Qiu et al., 2007; Ho and Yeh, 2009; Poort et al., 2012).

Similar to our approach, several other studies also make use of recurrent connections between different visual areas. Zwickel et al. (2007) studied the influence of feedback projections from higher areas in the dorsal visual stream on border-ownership coding. Feedback in their model is only to the border-ownership selective neurons, so they did not test their model on contour integration tasks. Domijan and Šetić (2008) proposed a model involving interactions between the dorsal and ventral visual streams for figure-ground assignment. In their model, there is no explicit computation of border ownership, but instead different surfaces are represented by different firing rates. Jehee et al. (2007) proposed a model of border-ownership coding involving higher visual areas, including areas TE and TEO. In their model, border-ownership assignment depends on the size of the figure, which is directly correlated to the specific level in the visual hierarchy of the model at which an object is grouped. They did not test their model on stimuli with noise, or study the effect of object-based attention. Sajda and Finkel (1995) proposed a complex neural architecture involving contour, surface, and depth modules that performs temporal binding through propagation of neural activity within and between populations of neurons. Tschechne and Neumann (2014) proposed a model quite similar to ours. Their model requires a repeated sequence of filtering, feedback, and center-surround normalization that

is solved in an iterative manner. Unlike in our model, V4 neurons in their model do not respond to straight, elongated contours. Also, their model only provides a coarse picture of the timing of contour integration and border-ownership assignment across visual areas. Layton et al. (2012) also proposed a model that performs border-ownership assignment. Their model introduces an additional neuron class ("R cells") that implements competition between grouping cells of different RF sizes (similar to a model by Ardila et al. (2012)), and the feedback to B cells is by means of shunting inhibition instead of the gain-modulation that we use in our model.

Previous experimental studies have suggested the involvement of multiple visual areas in contour integration and figure-ground segregation (Poort et al., 2012; Chen et al., 2014). The purpose of this chapter is to extend previous models of perceptual organization (Craft et al., 2007; Mihalas et al., 2011) to explain how feedback grouping circuitry can implement the mechanisms necessary to accomplish both contour integration and figure-ground assignment. Most models mentioned above reproduce results restricted to a single set of experiments (*e.g.* contour integration or figure-ground segregation). In contrast, our model is able to reproduce results from at least three sets of experiments using the same set of network parameters. Our model provides a general framework for understanding how features can be grouped into proto-objects useful for the perceptual organization of a visual scene. In addition, our model also allows us to explain effects of object-based attention and the role of feedback in parsing visual scenes, areas of research which have not been extensively

studied.

2.2 Methods

2.2.1 Model structure

The model consists of areas V1, V2, and V4 (Figure 2.1). Input comes from a binary-valued orientation map with four orientations ($0, \pi/4, \pi/2$, and $3\pi/4$ relative to the horizontal). The input signal is first represented in V1 and then propagated to V2 and V4 by feedforward connections. Area V4 provides feedback to lower areas (see Appendix A for equations). Neurons in higher areas have larger RFs and represent the image at a coarser resolution. Linear RF sizes in area V4 are four times larger than in V2, which, in turn, are twice as large as those in V1.

To achieve contour integration, we implement mutual excitatory lateral connections between V1 edge (E) cells with the same orientation preference. These connections are similar to the local association fields used in other models (Li, 1998; Piëch et al., 2013). Background suppression is carried out through a separate population of inhibitory (IE) cells. The input from V1 edge cells activates border-ownership (B) cells in V2. The B cells inherit their orientation preferences from their presynaptic E cells and for each orientation, there are two B cell populations with opposing side-of-figure preferences. A combination of lateral connections within V2 and feedback connections from V4 (described below) is used to generate border-ownership selectiv-

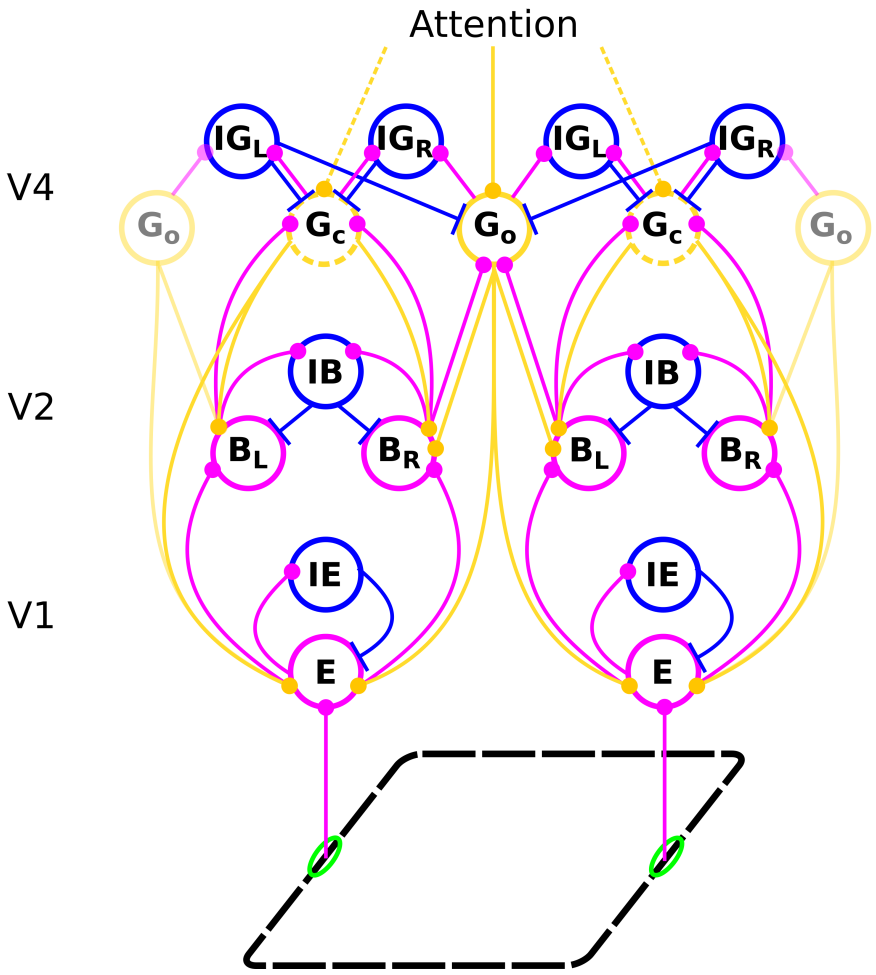


Figure 2.1: Structure of the model network. Each circle stands for a population of neurons with similar receptive fields and response properties. Magenta, blue, and orange lines represent feedforward excitatory, lateral inhibitory, and feedback excitatory projections, respectively. Top-down attention is modeled as input to the grouping cells and can therefore either be directed towards objects (solid lines) or contours (dashed lines) in the visual field (top).

ity. Inhibitory (IB) cells in V2 cause competition between B cells that have the same location and same orientation preference and opposite side-of-figure preference.

In V4, two different types of grouping cells exist. Contour grouping cells (G_c) integrate local edge information and are selective for oriented contours (Figure 2.2A). Object grouping cells (G_o) are sensitive to roughly co-circular arrangements of edges,

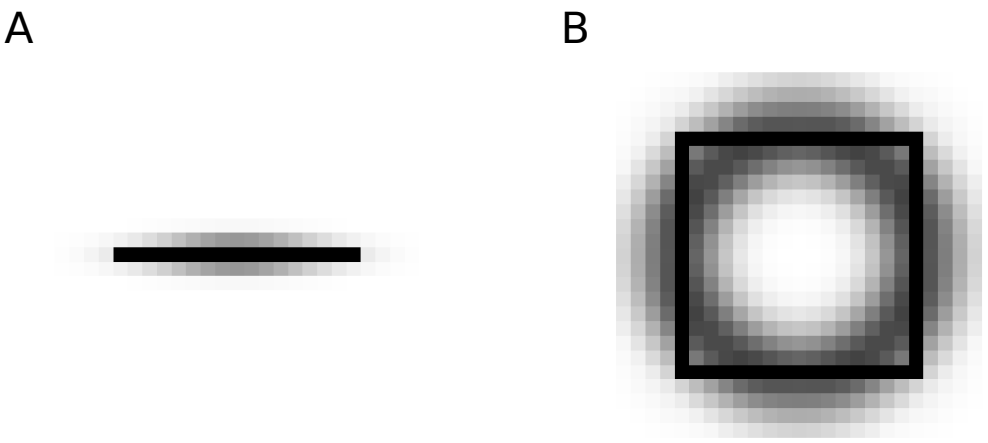


Figure 2.2: Spatial distribution of border-ownership cell to grouping cell connectivity; darker pixels indicate stronger connection weights. (A) Contour grouping neurons integrate features along oriented contours (horizontal line shown in black), emphasizing the Gestalt principle of good continuation. (B) Object grouping neurons integrate features in a co-circular pattern (square figure shown in black), emphasizing the Gestalt principles of convexity and proximity.

thus implementing Gestalt laws of good continuation, convexity of contour, and compact shape (Figure 2.2B). Competition between separate contours and objects is carried out by a population of inhibitory (IG) cells. Grouping (G_c and G_o) cells project back reciprocally to those B cells from which they receive input, and also to the E cells that project to those B cells. This feedback enhances the activity of E cells along contours and biases the competition between B cells to correctly assign border ownership along object boundaries. Importantly, feedback connections are modulatory, rather than driving, such that the feedback does not modify activities of cells that do not receive sensory input. Biophysically, this can be achieved if the feedback projections employ glutamatergic synapses of the NMDA type (Wagatsuma et al., 2016).

To model the effect of object-based attention, we assume that areas higher than V4 provide additional excitatory input to those grouping cells whose activity represents the presence of objects or contours in the visual scene, as shown in Figure 2.1. This attentional input is driving (as opposed to modulatory) but it is relatively weak; we select its strength as 7% of that of the driving input to the sensory (E) cells. In one part of this study (section 2.3.2), we model the effect of a lesion in V4 that removes the feedback completely by setting the weight of feedback connections from V4 to lower areas to zero.

Our approach is an extension of the proto-object-based model of perceptual organization proposed by Mihalas et al. (2011). Different from their approach, we include a new population of contour grouping neurons (the G_c cells) to explain recent results on cortico-cortical interactions during contour integration (Chen et al., 2014). As a result, top-down attention in our model can either be directed to contours or to objects. Our model is also able to reproduce the time course of neural responses in different visual areas, while the Mihalas et al. (2011) model only explains mean neural activities. In order to create more complex input stimuli, we also increased the number of model orientations from two to four. As a simplification to their model, we only include one scale of grouping neurons since we focus on mechanisms that do not require multiple scales.

2.2.2 Model implementation

Model neuronal populations (usually referred to as "neurons" in the following) are represented by their mean activity (rate coding). The activity is determined by a set of coupled, first-order nonlinear ordinary differential equations which was solved in MATLAB (MathWorks, Natick MA) using standard numerical integration methods. The mean firing rate is necessarily positive, therefore units are simple zero-threshold, linear neurons which receive excitatory and inhibitory current inputs with their dynamics described by,

$$\tau f'(t) = -f + \left[\sum W\right]_+ \tag{2.1}$$

where f represents the neuron's activity level and τ its time constant, chosen as $\tau = 10^{-2}$ s for all neurons. The sum is over all W which are the neuron's inputs, f' is the first derivative of f with respect to time, and $[\,]_+$ means half-wave rectification.

All simulations were performed on a 300-core CPU cluster running Rocks 6.2 (Sidewinder), a Linux distribution intended for high-performance computing. A total of 100 simulations were performed for each experimental condition, and our results are based on the mean neural activities averaged over these simulations with different randomly selected stimulus noise patterns, see sections 2.2.3 and 2.2.4.

To constrain our model parameters, we used three sets of neurophysiological data. The first comes from recent contour grouping results (Chen et al., 2014), and our

model was able to largely reproduce the magnitude and time course of contour integration at both the V1 and V4 levels. The second comes from studies of border-ownership coding (Zhou et al., 2000) and we show that our model not only explains the emergence of border-ownership selectivity but also its approximate time course as reported in that study. The third set of experimental data constraining our model is from the study by Qiu et al. (2007) of the interaction between border-ownership selectivity and selective attention. In order to fit the model parameters, we started from the parameters given by Mihalas et al. (2011), and modified them to fit the larger body of experimental results to include contour integration, border-ownership selectivity, and attentional selection.

2.2.3 Contour integration experiments

For the contour integration experiments reported by Chen et al. (2014), awake behaving monkeys were trained to perform a two-alternative forced-choice task using two simultaneously presented patterns, one containing a contour embedded in noise and one that was noise only (see Figure 2.3 for examples). The patterns were composed of 0.25° by 0.05° bars distributed in 0.5° by 0.5° grids. The diameter of the patterns was 4.5°, and the number of bars in the embedded contour was randomly set to 1, 3, 5, or 7 bars within a block of trials in order to control the saliency of the contour. To obtain a reward, the monkey had to saccade to the pattern that contained the contour. When the number of bars was set to 1, both presented stimuli

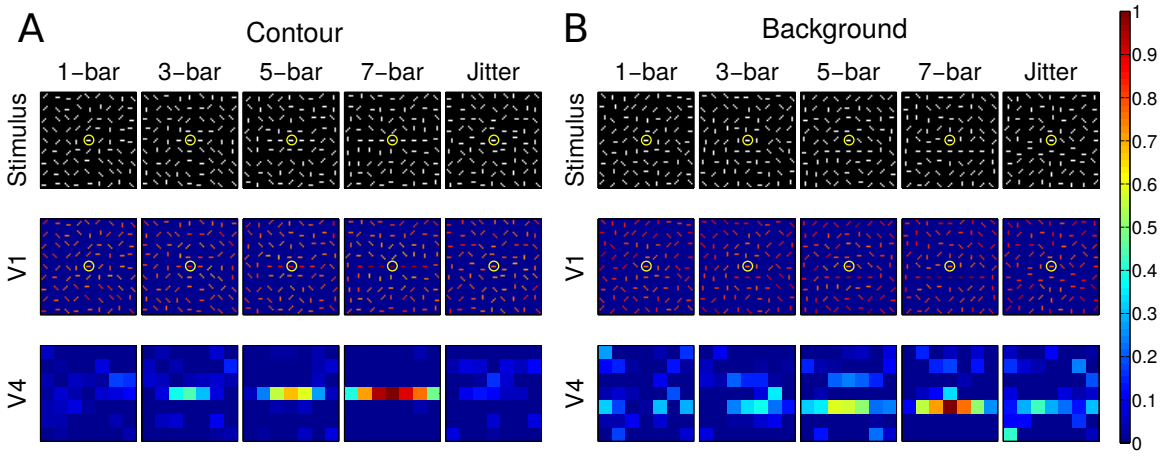


Figure 2.3: Normalized V1 E cell and V4 G_c cell population responses to contours of varying lengths in either the contour (A) or the background (B) condition. In (A), the "recorded" neuron is on the contour; in (B), it is offset from the contour. The top row shows stimuli, the second and third rows show activity in model areas V1 and V4, respectively. Yellow circles mark the RFs of the V1 neurons whose activity is shown in Figure 2.4. Columns in each condition show, from left to right, increasing contour length, with the right-most column showing a jittered stimulus configuration (see text). Neural activity is color coded and normalized to the 7-bar stimulus in both contour and background conditions, with warmer colors representing higher activity (see color bar at right).

were noise patterns, and the monkey was rewarded randomly with a probability of 50%. While the animals were performing the task, simultaneous single- or multi-unit recordings were made in area V1 and V4 neurons with overlapping receptive fields.

We modeled these experiments by creating visual stimuli contained in a 4.5° by 4.5° area. We "recorded" from the V1 receptive field that was at the center of the stimulus (marked by the yellow circles in Figure 2.3), as well as from the corresponding V4 neuron. Input from this area was projected onto a V1 layer of 64×64 neurons, each with a receptive field size of $\sim 0.7°$ so the receptive fields overlapped ten-fold at each location in the visual field. We divided the input field into a 9×9 grid

(each grid point at the center of a 0.5° by 0.5° area) and we placed at each grid location a stimulus bar consisting of three adjacent pixels, corresponding to a size $\sim 0.21°$ by $\sim 0.07°$. Each stimulus bar had one of four orientations, $0, \pi/4, \pi/2$, or $3\pi/4$. As in the Chen et al. (2014) experiment, contours consisted of 1, 3, 5, or 7 adjacent stimulus bars. We positioned the contour either at the center of the visual field so that the center element of the contour was in the RF of the "recorded" cell (Contour condition, Fig. 2.3A), or offset from the center of the visual field, such that the "recorded" neuron was *next* to the contour (Background condition, Fig. 2.3B). We changed the length of the contour by adding bars to both ends of the contour. Due to their size, V4 receptive fields basically enclosed the entire contour. For more details on the stimulus configuration, see Section 2.3.1.

2.2.4 Figure-ground segregation experiments

For the figure-ground segregation experiments (Zhou et al., 2000; Qiu et al., 2007; Zhang and von der Heydt, 2010), awake, behaving monkeys were trained on a fixation task. Receptive fields of each recorded neuron in areas V1 and V2 were first mapped to determine the optimal stimulus properties for that neuron. Afterwards, in some experiments, a square shape was presented on a uniform gray background with one edge of the square centered on the receptive field of the neuron at the neuron's preferred orientation. In other experiments (results shown in Figure 2.9) the stimulus consisted of two partially overlapping squares, and again the receptive field of the

recorded neuron was centered at its preferred orientation on one edge of one of the squares. The size of the square varied between experiments but it was always chosen such that the closest corner was far away from the classical receptive field of the recorded neuron. The square was presented in two positions between which it was "flipped" along the long axis of the neuron's receptive field. For instance, if the preferred orientation was vertical, the square was presented either to the left or the right of the cell's receptive field (we used this example in the choice of the indices in Figure 2.1). The difference in the firing rate of the neuron for when the square appears on one side versus the other side is defined as the border-ownership signal. Importantly, in all stimulus conditions, local contents within the receptive field of the neuron remained the same between these two conditions; only global context changed, so the neuron had to integrate information from outside its classical receptive field.

We modeled these experiments by creating visual stimuli that were projected onto the V1 layer. The input to the simulation was either a single square of a size that maximally activated G_o grouping cells of the size chosen in our model, or two partially overlapping squares, as shown in Figure 2.9, with each of these squares having the same optimal size. In other models (Craft et al., 2007; Mihalas et al., 2011; Russell et al., 2014), grouping cells of many scales are present, covering the range of possible objects in the input. We calculated border-ownership selectivity at the V2 level using the vector modulation index defined in section 2.2.5. In order to create noisy versions of the single square image (Figure 2.7), we followed a similar approach as in the

contour integration experiments, section 2.2.3. We again divided the visual field into a 9×9 grid and we positioned horizontal and vertical bars at specific grid points to generate a square. We then placed a stimulus bar at all other grid points, with their orientations randomly chosen from four possibilities, $0, \pi/4, \pi/2$, and $3\pi/4$.

2.2.5 Quantitative assessment of border-ownership selectivity: vector modulation index

While the border-ownership signal discussed in the previous section (the difference in firing rates between presentations of a figure on the two sides of a a neuron's RF) is useful to characterize border-ownership selectivity for that particular neuron, description of population behavior requires a more general measure. We use the vector modulation index, introduced by Craft et al. (2007) and defined by the expression

$$\overrightarrow{\mathbf{v}}(x, y) = m_{\hat{\imath}}(x, y)\hat{\imath} + m_{\hat{\jmath}}(x, y)\hat{\jmath} \tag{2.2}$$

where $\hat{\imath}$ and $\hat{\jmath}$ are unit vectors along the horizontal and vertical image axis, respectively, and the components $m_{\hat{\imath}}(x, y)$ and $m_{\hat{\jmath}}(x, y)$ are the usual modulation indices

along their respective axes, defined as

$$\begin{aligned} m_{\hat{\imath}}(x,y) &= \frac{\sum_\theta B_\theta(x,y)\cos\theta}{\sum_\theta |B_\theta(x,y)\cos\theta|} \\ m_{\hat{\jmath}}(x,y) &= \frac{\sum_\theta B_\theta(x,y)\sin\theta}{\sum_\theta |B_\theta(x,y)\sin\theta|} \end{aligned} \tag{2.3}$$

Here, $B_\theta(x,y)$ is the border-ownership signal (difference between the activities of the two opposing B neurons) at the preferred orientation θ at position (x,y), and θ runs over all angles taken into account in the model (four directed orientations in our case, namely $0, \pi/4, \pi/2$, and $3\pi/4$, each with both side-of-figure preferences). Vertical bars in the sums in the denominators indicate absolute values.

Both components in Eq. 2.3 are limited to values between +1 and -1. For the x-component, for instance, a positive value of $m_{\hat{\imath}}(x,y)$ signifies that the figure is to the right of position (x,y) and a negative value signifies that the figure is to the left. Its absolute value indicates the "strength" of the border-ownership signal, with zero being equivalent to ambivalence between left and right. Corresponding comments apply to the y-component, $m_{\hat{\jmath}}(x,y)$, regarding the figure's position upward or downward of (x,y). The direction of the vectorial modulation index $\overrightarrow{\mathbf{v}}(x,y)$ defined in Eq. 2.2 indicates the position of the foreground figure in the two-dimensional image plane relative to the point (x,y). For instance, positive values in both components $[m_{\hat{\imath}}(x,y) > 0,\ m_{\hat{\jmath}}(x,y) > 0]$ indicate that the figure is located upwards and to the right of (x,y).

2.3 Results

2.3.1 Contour enhancement in V1 and V4

We examined contour-related responses in our model using visual stimuli composed of collinear bars among randomly oriented bars (Figure 2.3), closely matching the stimuli used in the physiological experiments by Chen et al. (2014). The number of collinear bars constituting an embedded contour was set to either 1, 3, 5, or 7 bars, determining the length of the embedded contour, which also controlled its saliency. When the number of collinear bars was one, the stimulus was identical to a noise pattern. We compared the activity of model neurons whose RFs are centered on the contours or in the noise background (but close to contours) with that obtained in the analogous positions during neurophysiological recordings.

V1 responses were split into those of neurons on contour sites (C-sites) and background sites (B-sites). For contour sites (Figure 2.3A), the embedded contour was centered on the RF of a neuron with a preferred orientation matching that of the contour. For background sites, the contour was laterally placed 0.5° away from the RF center of the recorded neuron, and a background bar was placed in the RF (Figure 2.3B), with the contour orientation again matching the preferred orientation of the recorded neuron. Both V1 and V4 neurons along the contour showed increased activity with contour length. Neurons on the background showed increased suppression with contour length. Correspondingly, we show in Figure 2.4 the responses of neurons

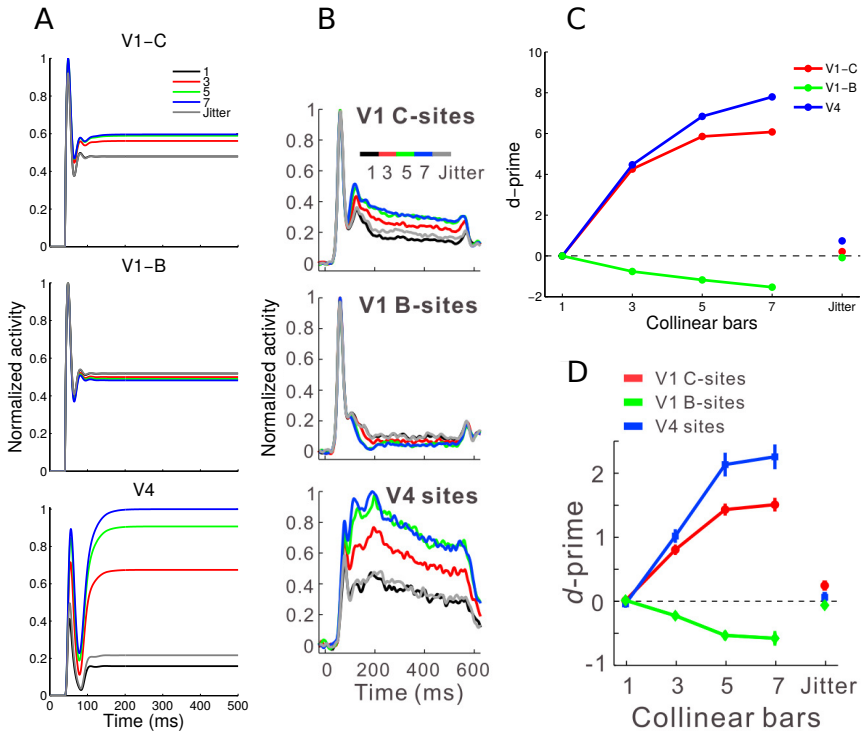


Figure 2.4: Normalized V1 E cell (contour and background sites) and V4 G_c cell neuronal activity and contour-response d' to contours of varying lengths. (A) V1 contour (top) and background (middle) sites and V4 sites (bottom). The jitter condition involved a 7-bar pattern where each bar was laterally offset to disrupt collinearity. (B) Corresponding experimental observations showing normalized and averaged PSTHs from the Chen et al. (2014) study. (C) Contour-response d' was higher for the V4 sites compared to the V1 contour sites. V1 background sites had increasingly negative d' with longer contours. (D) Corresponding experimental observations, showing the mean contour-response d' from the Chen et al. (2014) study. Panels B and D are modified from Figure 2 of Chen et al. (2014). All model results (neural responses and contour-response d') are averages for a single neuron over 100 simulations.

whose preferred orientations align with the contours (yellow circles in Figure 2.3).

Except for an input delay of 40ms corresponding to the duration of visual processing from the retina to V1, we did not add any time delays in the feedforward or feedback connections of our model, as we were not attempting to reproduce any specific latency effects. Nevertheless, our model generally reproduced the dynamics

of neural responses to contours in both V1 and V4 observed in the Chen et al. (2014) experiment (Figure 2.4A). The most salient feature of the neuronal responses is that the levels of sustained activity differ based on the number of bars in the embedded contour. This is observed both for contour sites and for background sites but, importantly, this effect went in opposite directions for these two cases. As the number of collinear bars increased from one to seven, V1 contour sites centered on the contour showed increased activity, with response saturation after five bars. In contrast, V1 background sites that were offset from the embedded contour showed a decrease in activity with increasing number of bars in the background (response suppression). Similar to V1 contour sites, V4 sites showed saturating responses with increasing contour lengths (Figure 2.4A, bottom). These results were qualitatively similar to those obtained in the Chen et al. (2014) experiments which are reproduced in Figure 2.4B (their Figure 2A).

The model data also showed strong onset transients in both (contour and background) V1 populations (Figure 2.4A, top and center), again in good agreement with experimental results (Figure 2.4B, top and center). Transients in V4 neurons were weaker, in both model and experimental data, and nearly absent in the experimental data (though not the model) for longer contours (Figure 2.4B, bottom). The transient peak observed in the model results is due to a sharp suppression of the activity level for a short (< 50ms) period which is not observed in the empirical data. We believe that this suppression is due to the strong inhibition at the V4 level between G and

IG cells, without equivalent excitation between different G cells.

Following Chen et al. (2014), we quantitatively analyzed the contour responses using the d-prime (d') metric from signal detection theory (Green and Swets, 1966), which quantifies the difference in distributions of mean neuronal firing rates between a contour pattern and the noise pattern integrated over the whole interval shown in Figure 2.4A,B, *i.e.* 0-500 ms. Neuronal responses to the 1-bar pattern (the noise pattern) were the baseline for examining contour-related responses in V1 and V4, this pattern therefore had a contour-response d' of zero. The contour-response d' increased with contour length for both the V1 contour and V4 sites, and d' decreased with contour length for the V1 background sites, Figure 2.4C.

The agreement between model (Figure 2.4C) and experimental results (Figure 2.4D) is striking. One difference we note is that the absolute values of all model d' substantially exceed the corresponding experimental values. This is to be expected since no noise was included in the model (other than the random orientation of input stimulus bars which is also present in the experimental approach) while there are surely multiple sources of noise in the biological system. We have not thoroughly investigated this question but it is highly likely that addition of noise to the model will decrease the d' values.

At first sight, it seems possible that V4 neurons respond with a higher firing rate to longer contours than to shorter ones simply as a consequence of the large size of RFs in V4. In this view, the enhanced responses in V4 with increasing contour length is

due to the spatial summation of many bars within the RF at the optimal orientation, independent of their precise location in the RF. To investigate this possiblity, Chen et al. (2014) introduced a "jitter" condition to the 7-bar contour, where alternating collinear bars were laterally offset by a small amount (much less than the receptive field size) in order to disrupt the collinearity of the original contour. They showed that jittering disrupted the contour integration process and reduced the neural responses in V1 and V4 close to baseline levels (Figure 2.4B, gray lines). We found the same result in our model, Figure 2.4A, gray lines. Furthermore, in the jitter condition, contour-response d' approached baseline for the V1 and V4 sites, as shown in the rightmost points for Figure 2.4D for experimental data and Figure 2.4C for model results. In both cases, no substantial difference between the jitter condition and the baseline noise condition was observed.

We also investigated the orientation and position dependence of contour-related responses in V1 and V4, and found close agreement of our model results with experimental data (Chen et al., 2014). Due to space constraints, these results are presented in Appendix A.

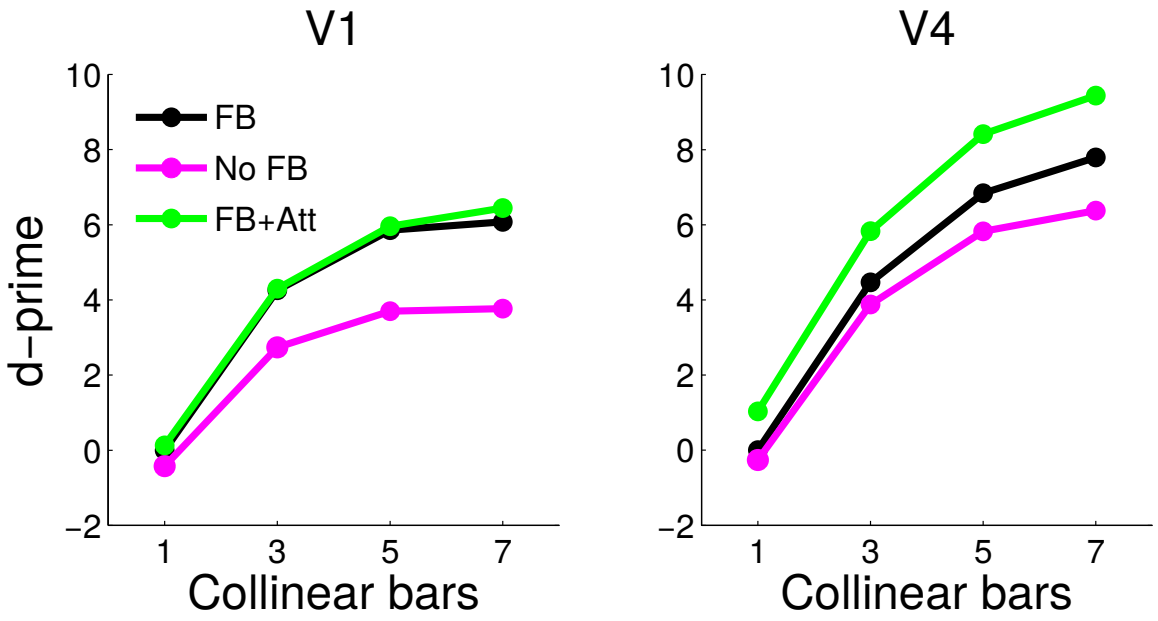


Figure 2.5: Contour-response d' in V1 E cells (A) and V4 G_c cells (B) for the model with (green) and without attention (black), and for the model with feedback removed (magenta). Attention strongly increased contour-response d' in V4 (B), while the lack of feedback strongly decreased contour-response d' in V1 (A).

2.3.2 The role of feedback and attention in contour grouping

While different forms of attention exist that can be flexibly used for different tasks, we choose here to focus only on the mechanisms of object-based attention (Egly et al., 1994; Scholl, 2001; Kimchi et al., 2007; Ho and Yeh, 2009). We postulate that attention to objects acts at the level of grouping neurons, in agreement with the Mihalas et al. (2011) model. As shown in that study, modulation of the activity of grouping cells bypasses the need for attentional control circuitry to have access to detailed object features. Instead, grouping neurons are used as "handles" of the

associated objects and it is in their interaction with feature-coding neurons that features are assigned to objects. One of the consequences of this mechanism is that the spatial resolution of attentional selection is coarser than visual resolution since the smallest "unit" of spatial attentional deployment is the size of the receptive (and projective) field of grouping cells, which is considerably larger than the receptive fields of feature-coding neurons at the same eccentricity. Behavioral results show strong evidence in favor of this coarse attentional resolution (Intriligator and Cavanagh, 2001).

In our model, attention can be directed to objects, including contours. For the contour integration experiments, this is implemented as a top-down input to all contour grouping neurons at the attended location, but, importantly, there was no direct input to the feature-coding edge (E) cells. As we have seen before (Figure 2.4), d' in V1 and V4 populations increases with the number of collinear bars, even in the absence of attention. In addition, we now show that attention increases the contour-response d' for both V1 and V4 neurons, with the additive effect being much larger in V4 than in V1 (Figure 2.5). For the 7-bar contours, attention increases the contour-response d' in V1 from 6.08 to 6.66 and in V4 from 7.80 to 11.59. This is consistent with findings that attention has a much larger effect on higher-level neurons compared to early sensory neurons (review: Treue, 2001).

One of the advantages of computational modeling is that it allows the study of scenarios that are difficult to implement empirically. One question that is difficult

to answer experimentally is how removal of feedback from V4 to lower visual areas would change neuronal responses in V1 and V4, structures that are known to be connected bidirectionally (Zeki, 1978; Ungerleider et al., 2007). While it is possible to study the influence of feedback on V2 (and other areas) by cooling (Hupé et al., 1998) or pharmacological inactivation (Jansen-Amorim et al., 2012) of area V4, such manipulations only allow limited control of the effect on multiple brain structures. In contrast, a computational model allows the study of interactions between different areas with perfect control. In the model, we can eliminate all feedback from area V4 simply by "lesioning" the connections from the grouping neurons to the V2 and V1 neurons (setting their strength to zero), thus turning the model into a feedforward network. We found that this has the opposite effect of applying attention on the d' metric: the decrease in contour response d' was much larger in V1 compared to V4. Removing feedback reduced the 7-bar contour-response d' in V1 from 6.08 to 3.77, and in V4 from 7.80 to 6.38 (Fig. 2.5). We note that the contour-response d' in V1 is above zero even without feedback from grouping neurons because of the contribution of local excitatory connections to contour integration. This asymmetric effect in contour response d' in the two areas (V1 and V4) may point to the different roles of feedforward and feedback processing in early vision. We are not aware of any experimental manipulations that completely remove feedback from area V4 without changing the circuitry in other ways, so our results are a prediction awaiting experimental falsification.

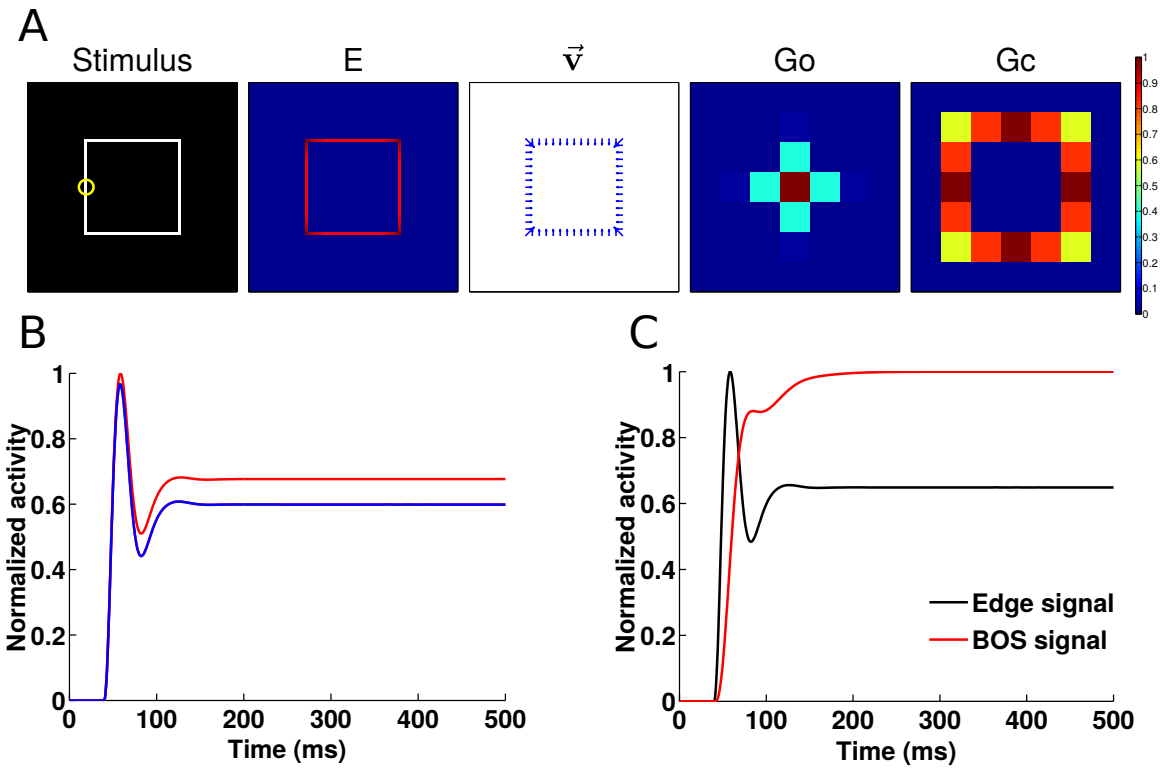


Figure 2.6: Figure-ground segregation of a square object as in the Zhou et al. (2000) experiments. (A) Shown left to right are the input stimulus, the edge cell activity (E), the border-ownership assignment along edges (shown as the vector modulation index $\vec{\mathbf{v}}$, section 2.2.5), the object grouping neuron activity (G_o) and the contour grouping activity (G_c). Activities are normalized within each map, and warmer colors indicate higher activity (see color bar at right). (B) Time course of normalized border-ownership cell activity for the preferred side-of-figure (red) and non-preferred side-of-figure (blue) for the receptive field marked by the yellow circle in panel A. (C) Timing of the normalized border-ownership signal (red) and the edge signal (black). The curves are normalized to the same scale (0–1) to show the time course of the responses.

2.3.3 Border-ownership assignment and highlighting figures in noise

We next apply our model to understanding border-ownership assignment, discussed in Section 2.2.4. We focus first on the standard square figure frequently used

in neurophysiological studies of this function (Zhou et al., 2000; Qiu et al., 2007; Sugihara et al., 2011; Williford and von der Heydt, 2013, 2014; Martin and von der Heydt, 2015). In Figure 2.6A, we show the input stimulus, the edge cell activity from area V1 of our model, the border-ownership vector modulation field from V2 (defined in Section 2.2.5), and the object and contour grouping cell activity from V4. Our model enhances V1 activity along the edges of the square, correctly assigns border ownership in V2 neurons (in agreement with Zhou et al., 2000), and enhances activity of V4 neurons at the center of the square and the edges of the square (object and contour grouping neurons, respectively).

We also show the time course of the border-ownership signal for a RF located along the left edge of the figure (indicated by the yellow circle in Figure 2.6A). The firing rate of a border-ownership selective neuron depends on which side the figure is presented on with respect to its RF. In Figure 2.6B, the preferred neuron has a side-of-figure preference to the right (red), while the non-preferred neuron has a side-of-figure preference to the left (blue). The firing rate difference is the border-ownership signal, whose steady-state value is used to compute the vector modulation index $\overrightarrow{\mathbf{v}}$ (Section 2.2.5). In Figure 2.6C, we show that the border-ownership signal (red) appears rapidly and with short latency compared to the onset of the edge signal (black). Experimental data show that the border-ownership signal appears ˜20 ms after the onset of edge responses (Zhou et al., 2000). In our model, the border-ownership signals appears almost instantly. We believe that in addition to

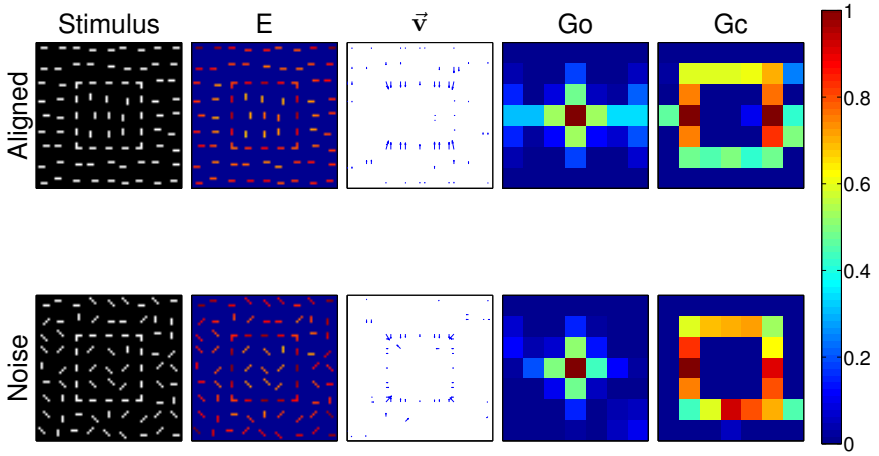


Figure 2.7: Figure-ground segregation of a square object with aligned contour elements (top row) and with noise contour elements (top row). Shown are (left to right) the input stimulus, the edge cell activity (E), the border-ownership assignment along edges (shown as the vector modulation index $\vec{\mathbf{v}}$, section 2.2.5), the object grouping neuron activity (G_o) and the contour grouping activity (G_c). Activities are normalized within each map, and warmer colors indicate higher activity (see color bar at right).

visual input, border-ownership cells also receive grouping feedback and that this takes additional time which is not included in our model (see Craft et al., 2007, as an example of a model with latencies).

We then extend this approach by adding additional oriented bars to the stimulus, see Figure 2.7. In the top row of the figure, results are shown when the bar orientation within the figure differs from that of the background, similar to the texture-defined figures used in the Lamme (1995) experiments. In the bottom row, bars within and outside the figure are all oriented randomly, similar to the noise stimuli used in the Chen et al. (2014) study. As in Figure 2.6, we again show the responses of different populations of neurons in our model. For both types of stimuli, the edges of the square, even when broken up into different bars, are still enhanced while the

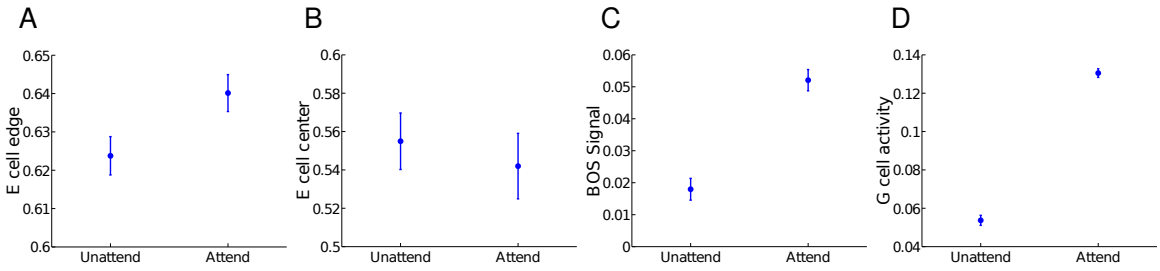


Figure 2.8: Attentional modulation of different neuronal populations aids figure-ground segregation in the presence of noise. Each panel (A-D) shows the means and standard deviations of neural activity with and without added attention over 100 different simulations. (A) Edge cell activity along the contour of the object shows enhancement with attention. (B) Edge cell activity in the center of the figure shows suppression with attention. (C) The BOS signal along the contour of the object shows enhancement with attention. (D) Object grouping cell (Go) activity centered on the object shows enhancement with attention.

background bars are suppressed, especially within the square. Interestingly, for the aligned stimulus (top row), only the top and bottom edges of the square show border ownership modulation. This may be an artifact of our modeling procedure, where the edges of our figure are defined by contour elements instead of texture discontinuities. For the noise stimuli (bottom row), border ownership is still assigned correctly along the edge of the square, but the noise results in occasional nonzero border-ownership cell activity at other points in the image as well. For both types of stimuli, grouping cell activity is still centered on the square, and contour grouping neurons still highlight the edges of the squares, but there is also noticeably increased noise.

2.3.4 Interaction between border-ownership assignment and attention

Although border-ownership selectivity emerges independently of attention (Qiu et al., 2007), attention may help to facilitate figure-ground segregation in the presence of noise. For the square figure with noise, we found that in our model, attention increases the responses of neurons along the edge of the figure (unpaired t-test, $p = 2.32 \times 10^{-59}$) and suppresses those in the center (unpaired t-test, $p = 3.42 \times 10^{-8}$). In addition, attention increases border-ownership modulation along the edge of the figure (unpaired t-test, $p = 1.37 \times 10^{-143}$) and increases the activity of object grouping neurons in the center of the figure (unpaired t-test, $p = 1.53 \times 10^{-239}$). All effects were small but highly significant, and were based on the differences in summed activity of neurons along the edge or center of the figure over a total of 100 simulations. Figure 2.8 shows the average neural activity in the different populations of neurons in our model, with and without the effect of attention.

Attention must also operate in cluttered environments where multiple objects may be present. Qiu et al. (2007) studied border-ownership responses in area V2 when two overlapping squares were presented and attention was directed either to the foreground or background square. Mihalas et al. (2011), using a model closely related to ours but without G_c cells, reproduced the experimental finding that border-ownership modulation was strong when attention was on the foreground figure but weak when

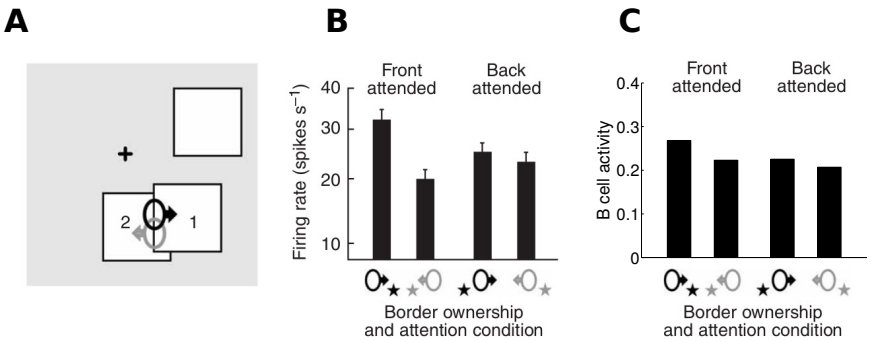


Figure 2.9: Quantitative comparison of model performance to neurophysiological findings (Qiu et al., 2007) for border-ownership coding of overlapping figures. (A) The stimulus configurations used are shown, with neurons coding right border ownership (black) and left border ownership (gray) when attention was focused on either foreground square 1 (front attended) or on background square 2 (back attended). (B) The responses of border-ownership selective cells recorded in V2 are shown: bars indicate the average firing rate for each stimulus condition. (C) Model B cell responses to analogous stimulus conditions. For both the model and the experiments, border-ownership modulation was strong when attention was on foreground but weak when attention was on background. Panels A and B are modified from Figure 3 of Qiu et al. (2007).

attention was on the background figure. Our model reproduces this finding, see Supplementary Figure A.3. The quantitative effect on border-ownership selectivity is shown in Figure 2.9. Our model also reproduces the experimental finding that border-ownership modulation is strong when attention is on the foreground figure (Figure 2.9C, "Front attended") and weak when attention is on the background figure (Figure 2.9C, "Back attended").

These results demonstrate that the object and contour grouping neurons are able to assist with early level segmentation of objects in noise and clutter. Previous experimental studies have only tested squares without noise, although the effect of figures defined by broken contours has been investigated before (Zhang and von der

Heydt, 2010). Our results predict that border-ownership assignment and grouping are robust even in the presence of noise, clutter, and interruptions in figure borders, and that attention may further aid this process.

2.4 Discussion

2.4.1 Model predictions

Our model predicts that attentional modulation is specific to the attended contour or object, rather than being defined purely spatially. This is possible because, in our model, attention modifies firing rates of grouping cells, rather than of elementary feature-coding cells. Attending to the contour increases contour-response d' in both V1 and V4, consistent with experimental results showing increased contour-related responses after animals had been trained to perform a contour detection task, compared to when they performed a separate set of tasks in which the contour was behaviorally irrelevant (Li et al., 2008). We note that Li et al. (2008) only studied neural responses in V1, while our model makes predictions about attention-related changes to contour-response d' in both V1 and V4. We also find that attention modulates border-ownership activity in V2 in an object-based manner. As a result, our model makes predictions about neural activity in visual areas V1, V2, and V4 across different stimuli and tasks.

We predict that the interaction of modulatory feedback from grouping cells with

local inhibition enhances the representation of the figure and suppresses the representation of the background. Indeed, our model produces background suppression for isolated contours as well as for figures embedded in noise. We note that this prediction is different from what others have observed in texture-defined figures (Lamme, 1995; Lee et al., 1998), where there is generally response enhancement of the center of the figure. This difference may be due to how the figure is defined, by its contour in our experiment and by texture in Lamme (1995) and Lee et al. (1998). It is possible that if feedback from object grouping neurons is to the center of the figure instead of its edges, we may observe enhancement of activity at the center of the figure. Understanding how border-ownership assignment interacts with filling-in of surfaces is a future direction of research. Furthermore we predict that attention, in addition to enhancing the figure in an object-based manner, also helps to suppress noise in the background.

We also predict that removing feedback from V4 to lower visual areas reduces neural responses in V1, while having a smaller effect in V4. The activity of V4 neurons is also affected due to the recurrency of the network model. This could be tested experimentally by the same contour detection task used by Chen et al. (2014), and measuring the contour-response d' in V1 and V4 from reversible inactivation of feedback connections. Although complete and selective deactivation of V4 feedback to areas V1 and/or V2 is technically challenging, there have been attempts to study the effect of this type of feedback either through extra-striate lesions (Supèr and Lamme,

2007) or reversible inactivation (Jansen-Amorim et al., 2012). Anesthesia presumably also decreases top-down influences, and indeed reduces contour-related V1 responses (Li et al., 2008) and figure-ground segregation (Lamme et al., 1998).

2.4.2 Comparison to other models

Many have argued that contour integration and figure-ground segregation are largely local phenomena that rely on lateral connections (Grossberg, 1994, 1997; Li, 1998; Zhaoping, 2005; Piëch et al., 2013). While some of these models include a role for top-down influences (Li, 1998; Piëch et al., 2013), they do not offer a specific mechanism by which higher visual areas representing object-level information selectively feed back to lower visual areas containing feature-level information about the object. In contrast, our model is explicit in that feedback connections from higher visual areas modulate the responses of early feature-selective neurons involved in the related processes of contour integration and figure-ground segregation. Our model thus is a member of a broad class of theoretical models that achieve image understanding through bottom-up and top-down recurrent processing (Ullman, 1984; Hochstein and Ahissar, 2002; Roelfsema, 2006; Epshtein et al., 2008). In comparison to similar models, our model is able to reproduce experimental findings from two traditionally separate fields of study– contour integration and figure-ground assignment. Importantly, using the same set of network parameters, the grouping cells in our model are able to represent proto-objects (both contours and extended objects) and provide a

perceptual organization of the scene. We show that this perceptual organization is critical for interfacing with top-down attention, and that it provides a general theoretical framework for understanding how feedback connections and an hierarchy of visual areas can be used to group together the features of an object.

2.4.3 Roles of V1 and V4 in visual processing

While V1 neurons have small RFs that accurately code for orientation, V1 also shows strong background inhibition off the contour. This property allows V1 neurons to enhance contours and suppress noise in the image at a high spatial resolution. V4 neurons, on the other hand, have large RFs that integrate local feature information over large areas of visual space and provide a coarse, proto-object representation of contours and objects. Feedback from V4 can then be refined by the lateral connections present within early visual areas, which aids in the enhancement of the figure and its edges in the image.

Even when V1 neurons receive no feedback from V4, there is an increase in contour-response d' with contour length (Figure 2.5). This contour facilitation is solely due to the excitatory lateral connections present in V1, and is weaker without feedback. As a result, feedback from V4 may not be necessary for contour facilitation, but it interacts with local lateral connections in a push-pull manner – neurons along the contour are enhanced, while elements on the background are suppressed. Not surprisingly, removing this type of feedback has a larger effect on V1 neurons

compared to V4 neurons, although the activity of V4 neurons is also affected due to the recurrency of the network model.

2.4.4 Contour and object grouping neurons

Contour grouping neurons have direct experimental support through the recent neurophysiological experiments published by Chen et al. (2014). In our model, relatively few numbers of grouping neurons (both contour and object) are required. The spatial resolution of the grouping process does not need to be very high, as grouping cells only provide a coarse template of the contours and objects present in the scene. Assuming that the activity of grouping cells represents proto-objects with a similar resolution as attention, the total number of grouping cells may be less than 2% of the number of border-ownership cells (Craft et al., 2007). In the contour integration experiments (Chen et al., 2014), many V4 neurons were found to respond to straight, elongated contours, while in our model, only a subset of the grouping neurons are selective for contours. One possible explanation for this is overtraining – monkeys performed the task a very large number of times each day for many months, and neural plasticity may have generated many V4 neurons which respond to contours.

There is no unambiguous neurophysiological evidence for object grouping neurons yet, although previous studies have found neurons in V4 that respond to contour segments of various curvatures (Pasupathy and Connor, 2002; Brincat and Connor, 2004). The receptive fields of these neurons are similar to those proposed by Craft

et al. (2007). Other types of grouping neurons may also exist, including those that respond to gratings (Hegdé and Van Essen, 2007), illusory surfaces (Cox et al., 2013), or 3D surfaces (He and Nakayama, 1995; Hu et al., 2015). We do not attempt to model the whole array of grouping neurons that may exist, but only those necessary for reproducing the neurophysiological experiments referred to here.

In our model, orientation-independent attentional input to contour grouping cells was used to enhance neural responses to a contour at a given location. We note that this form of attention is analogous to the size- and location tolerant attentional selection process proposed by Mihalas et al. (2011). The local circuitry of their model was able to sharpen a relatively broad and nonspecific attentional input to match the size and location of a figure in the visual scene. Similarly, the local circuitry of our model transforms the orientation-independent attentional input such that it only enhances the contour of the correct orientation. We note that attention also has a suppressive effect in our model, essentially inhibiting unattended objects and locations. There is physiological support for this mechanism (Wegener et al., 2004; Hopf et al., 2006; Sundberg et al., 2009; Tsotsos, 2011). Furthermore, previous results show suppression of border-ownership activity along the shared edge of overlapping squares when the back square was attended, but not the front square (Qiu et al., 2007). Finally, psychophysical experiments demonstrate an "object superiority effect," where reaction times are fastest when attention is directed to targets that are part of the cued object, and slowest when targets are outside of an object (Egly et al., 1994;

Kimchi et al., 2007).

Complementarily, feature-based attention acts broadly across the visual scene and increases the responses of all components that share similar feature attributes (*e.g.* color, orientation, or direction of movement) with the attended component (Motter, 1994; Treue, 1999). Orientation-specific forms of attention can enhance neural responses in V1 and V4, but do not significantly alter tuning curves or selectivity (McAdams and Maunsell, 1999). Our model may be able to reproduce similar results by essentially changing the form of the attentional input to be orientation-specific, *i.e.* top-down attention targets a *single* population of contour grouping neurons with the same orientation preference. We expect that both object-based and feature-based forms of attention exist and can be flexibly used for different tasks.

2.5 Conclusion

Our model seeks to reproduce three different sets of experimental results, while making testable predictions for future experiments. We only included one scale of grouping neurons for simplicity, although multiple scales of grouping neurons are needed to account for the diversity in the scale of objects in the real world. Our model also assigns distinct roles to the different visual areas, edge processing in V1, border-ownership assignment in V2, and grouping of contours and objects in V4. However, the physiological properties of neurons in early visual areas have not been

fully characterized, and neurons in these different areas may have additional ranges of selectivity than the ones we assign them in our model. Finally, our model operates on artificial images composed of simple shapes such as contours or square figures. In order to truly understand grouping mechanisms in natural vision, our model must also be able to operate on natural images as input, where the number of potential objects and features are much richer. We propose such a model in the next chapter.

Chapter 3

A recurrent neural network for figure-ground organization of natural scenes

3.1 Introduction

Figure-ground organization is critical for visual understanding of the world around us. The process typically involves some form of segmentation, or dividing the input image into regions corresponding to objects and background. The border between any two regions is usually "owned" by a single object, and the correct assignment of each border to its corresponding object is thought to be a precursor to scene understanding. However, this task is difficult due to the clutter, occlusion, and wide variety

of features present in natural scenes. This problem has long fascinated researchers, including those from the fields of psychology (Wertheimer, 1923; Koffka, 1935), neuroscience (Zhou et al., 2000; Craft et al., 2007), and computer vision (Sajda and Finkel, 1995; Ren et al., 2006; Teo et al., 2015; Wang and Yuille, 2016). Despite this long line of research, our understanding of the neural basis of figure-ground organization remains surprisingly limited.

A key insight from neuroscience was the discovery of border-ownership cells. Border-ownership selectivity is a property of a majority of the neurons in V2, which encodes where an object is located relative to the neuron's receptive field (Zhou et al., 2000). When the edge of an object is presented in its receptive field (RF), a border-ownership cell will respond with different firing rates depending on where the object is located. For the example cell shown in Figure 3.1A, objects placed to the upper right of the cell's RF (stimuli outlined in red) cause the cell to fire at a higher firing rate compared to objects placed to the lower left of the cell's RF (stimuli outlined in blue). Importantly, the local stimulus within the cell's RF is identical in both cases. Border-ownership coding has been studied using a wide variety of artificial stimuli, including those defined by luminance (Zhou et al., 2000), motion (von der Heydt et al., 2003), disparity (Qiu and von der Heydt, 2005), and transparency (Qiu and von der Heydt, 2007), and more recently, natural stimuli such as faces (Hesse and Tsao, 2016; Ko and von der Heydt, 2017) and complex natural scenes (Williford and von der Heydt, 2016).

To explain this phenomenon, some computational models assume that border-ownership coding is achieved purely by feedforward mechanisms, such as the asymmetric organization of surrounds (Nishimura and Sakai, 2004, 2005; Sakai et al., 2012) or global surround inhibition (Supèr et al., 2010). Pure feedforward models predict similar latencies of the border-ownership signal regardless of the stimulus, but recent results show that border-ownership assignment of stimuli with illusory contours is delayed by $\sim$ 30ms compared to full stimuli (Hesse and Tsao, 2016). Other models propose propagation of neural activity along horizontal connections within early visual areas using a diffusion-like process (Grossberg, 1994; Sajda and Finkel, 1995; Baek and Sajda, 2005; Kikuchi and Akashi, 2001; Pao et al., 1999; Zhaoping, 2005; Zucker, 2012). However, these models have difficulties explaining the fast establishment of border ownership which appears about 25ms after the first stimulus response (Zhou et al., 2000). Propagation along horizontal fibers over the distances used in the experiments would imply a delay of at least $\sim$ 70ms (Girard et al., 2001, based on the conduction velocity of horizontal fibers in primate V1 cortex; we are not aware of corresponding data for V2). Furthermore, such models are difficult to reconcile with the observation that the time course of border-ownership coding is largely independent of figure size (Sugihara et al., 2011).

An alternative computational model involves populations of grouping ($\mathcal{G}$) cells which explicitly represent (in their firing rates) the perceptual organization of the visual scene (Craft et al., 2007; Mihalas et al., 2011). These cells are reciprocally

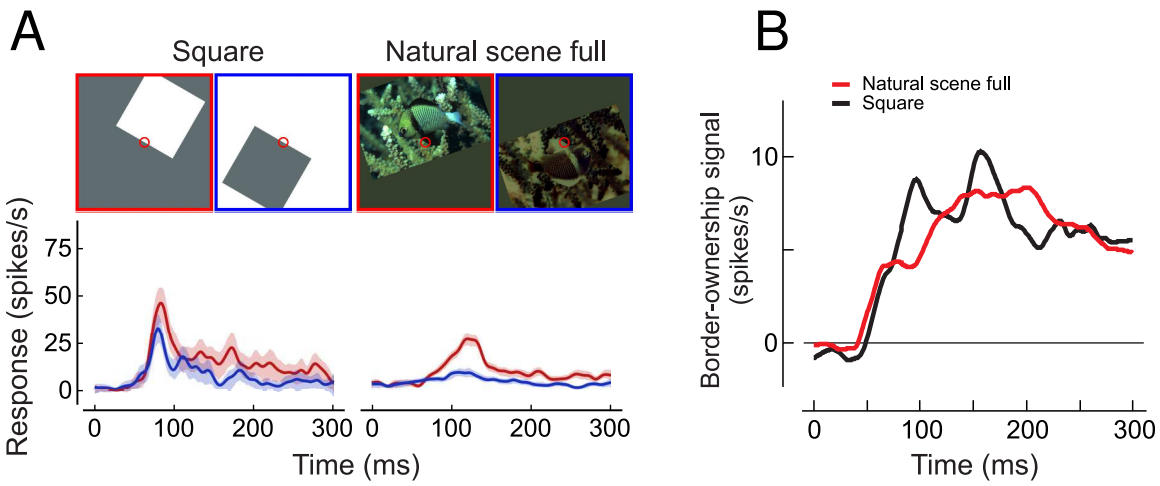


Figure 3.1: Consistency of border-ownership coding. (A) Border-ownership coding for an example cell. The cell has a preference for objects located to the upper right of its receptive field (RF) on both the square and natural scene stimuli, as indicated by higher firing rates. Red circles indicate the size and location of the cell's RF. Stimuli with objects to the upper right of the cell's RF ("preferred" side) are outlined in red, while stimuli with objects to the lower left of the cell's RF ("non-preferred" side) are outlined in blue. The cell's postimulus time histogram (PSTH) for the preferred side is shown by the red trace, while the PSTH for the non-preferred side is shown by the blue trace. Shading indicates 95% confidence intervals.

connected to border-ownership selective ($\mathcal{B}$) cells through feedforward and feedback connections. The combined activation of grouping cells and cells signaling local features represents the presence of a "proto-object" (we borrow this term from the perception literature; Rensink, 2000), resulting in a structured perceptual organization of the scene. This proto-object based approach is consistent with psychophysical and neurophysiological studies (*e.g.* Duncan, 1984; Egly et al., 1994; Scholl, 2001; Kimchi et al., 2007; Qiu et al., 2007; Ho and Yeh, 2009; Poort et al., 2012). A feedforward version of this model has been applied to natural scenes, where it outperformed other models in predicting the locations of human eye fixations (Russell et al., 2014).

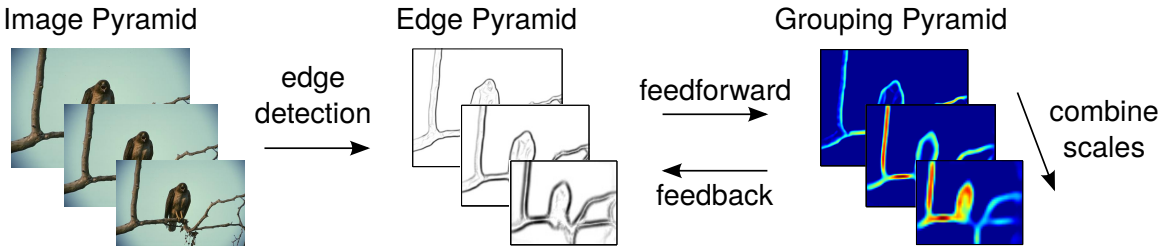


Figure 3.2: Overview of the model. The image is first successively downsampled in half-octaves to create an image pyramid (only three scales are shown). The same set of feedforward and feedback grouping operations is then applied at each level of the pyramid to achieve scale invariance. Feedback from grouping cells is combined across scales so that global context information can influence figure-ground segmentation. The model is run for a total of 10 iterations (one iteration includes a feedforward and a feedback pass through the model), and our final results are based on neural activity from the highest resolution scale of the image pyramid.

A substantial fraction of neurons show consistent border-ownership coding across natural scenes that matches their preference on artificial stimuli, with the timing of border-ownership signals being similar for both types of stimuli (Figure 3.1B). However, with the exception of a single computational model (Sakai et al., 2012), we are not aware of any models that have quantitatively tested border-ownership selectivity on natural scenes. Here, we propose a model based on recurrent connections that is able to explain border-ownership coding in natural scenes. We compare our model results with experimental results, and find a good match both in the timing of the border-ownership signals as well as the consistency of border-ownership coding across scenes. In order to compare our model with other computer vision approaches, we also benchmarked our model on a standard contour detection and figure-ground assignment dataset (Martin et al., 2001) and achieve performance comparable to these other approaches.

3.2 Methods

3.2.1 Model Structure

Our approach is an extension of the proto-object based model of saliency proposed by Russell et al. (2014) and includes recurrent connections for figure-ground assignment. At the core of our model is a grouping mechanism which estimates figure-ground assignment within the input image using proto-objects of varying spatial scales and feature types. These proto-objects provide a coarse segmentation of the image through perceptual organization of the scene into regions corresponding to objects and background.

To achieve scale invariance, the algorithm successively downsamples the input image in steps of $\sqrt{2}$ to form an image pyramid spanning five octaves (Figure 3.2). The kth level of the pyramid is denoted using the superscript k. Unless explicitly stated, any operation applied to the pyramid is applied independently to each level and each feature type. Each layer of the network represents neural activity, which can either be propagated from one layer to another via feedforward or feedback connections since our model is recurrent. We use a filter-based approach, where the receptive fields of neurons are described by kernels and the correlation operation is used to calculate the neural response to an input. The model was implemented using MATLAB (Mathworks, Natick, MA, USA).

The first stage of the model extracts edges from the input image based on either

luminance or color information (Figure 3.2). We use the combination of receptive fields (CORF) operator, which is a model of V1 simple cells with push-pull inhibition (Azzopardi et al., 2014). We chose the CORF operator due to its texture suppression properties, which can be beneficial when applied to natural images. Our model does not require a specific edge detection method and could be modified to use other front-end edge detectors (*e.g.* Gabor filters). In the following, we only describe model computations on the luminance channel, but the exact same computations are also performed on the two color channels (*i.e.* red-green and blue-yellow). As in Russell et al. (2014), the color channels were computed according to the methods outlined in the original Itti et al. (1998) visual saliency model.

For a given scale k, the output of the edge detection stage of the model are simple ($\mathcal{S}$) cells of eight different orientations θ and two contrast polarities, termed $\mathcal{S}^k_{\theta,L}(x,y)$ and $\mathcal{S}^k_{\theta,D}(x,y)$ (*i.e.* for light-dark edges L and dark-light edges D). For the two color channels, the edge polarities are determined by color-opponent responses (*e.g.* red-green edges and green-red edges). Only the signal strength at the optimal orientation at each spatial location was used as input to the network. This simplification significantly reduces computation time by eliminating the calculation of responses for the non-optimum orientations.

In contrast to previous approaches which combine simple cell responses into a contrast-invariant complex cell response (Russell et al., 2014), we keep the contrast-sensitive $\mathcal{S}$ cell responses as they provide an informative cue for grouping along object

edges. Objects tend to maintain similar contrast polarity along their boundaries, which may be useful for accurately determining figure-ground relationships. As a result, we have two sets of responses at each layer of our network corresponding to the two different types of contrast polarity.

Next, for a given angle θ, each $\mathcal{S}$ cell feeds into an opposing pair of border-ownership ($\mathcal{B}$) cells. As a result, $\mathcal{B}$ cells are also sensitive to contrast polarity, which is consistent with experimental findings (Zhou et al., 2000). For each contrast polarity, we used one-to-one connections between $\mathcal{S}$ cells of one orientation and the corresponding pair of $\mathcal{B}$ cells with the same preferred orientation, but opposing side-of-figure preferences.

$$\begin{aligned} \mathcal{B}^k_{\theta,L} &= \mathcal{S}^k_{\theta,L}(x,y) \\ \mathcal{B}^k_{\theta+\pi,L} &= \mathcal{S}^k_{\theta,L}(x,y) \end{aligned} \tag{3.1}$$

$$\begin{aligned} \mathcal{B}^k_{\theta,D} &= \mathcal{S}^k_{\theta,D}(x,y) \\ \mathcal{B}^k_{\theta+\pi,D} &= \mathcal{S}^k_{\theta,D}(x,y) \end{aligned} \tag{3.2}$$

To infer whether the edges in $\mathcal{B}^k_{\theta,L}(x,y)$ and $\mathcal{B}^k_{\theta,D}(x,y)$ belong to figure or ground, knowledge of proto-objects in the scene is required. This context information is retrieved from a grouping mechanism (Figure 3.2). Grouping cells ($\mathcal{G}$) integrate information from $\mathcal{B}$ cells, and either respond to light objects on dark backgrounds,

$\mathcal{G}_L^k(x,y)$, or dark objects on light backgrounds, $\mathcal{G}_D^k(x,y)$. This computation is similar to the use of center-surround cells in a feedforward version of the model (Russell et al., 2014). In contrast to their approach, our model does not require an additional class of center-surround cells, but instead allows $\mathcal{G}$ cells to directly integrate local feature information from $\mathcal{B}$ cells and then bias the activity of these same cells using reciprocal feedback connections. Our model runs in an iterative manner, with one iteration corresponding to a feedforward and feedback pass through the model. $\mathcal{G}$ cell activity is combined across scales before each feedback pass, which allows the model to more accurately determine figure-ground assignment in a scale-invariant manner (Figure 3.2).

A more detailed view of the structure of our model is shown in Figure 3.3. $\mathcal{G}$ cells integrate the $\mathcal{B}$ cell activity in an annular fashion. This allows $\mathcal{G}$ cells to show preference for objects whose borders exhibit the Gestalt principles of continuity and proximity. $\mathcal{G}$ cell activity is defined according to

$$\mathcal{G}_L^k(x,y) = \left\lfloor \sum_\theta [\mathcal{B}_{\theta,L}^k(x,y) - \mathcal{B}_{\theta+\pi,L}^k(x,y)] * v_\theta(x,y) \right\rfloor \tag{3.3}$$

$$\mathcal{G}_D^k(x,y) = \left\lfloor \sum_\theta [\mathcal{B}_{\theta,D}^k(x,y) - \mathcal{B}_{\theta+\pi,D}^k(x,y)] * v_\theta(x,y) \right\rfloor \tag{3.4}$$

where where $\lfloor \cdot \rfloor$ is a half-wave rectification, and $*$ is the correlation operator defined

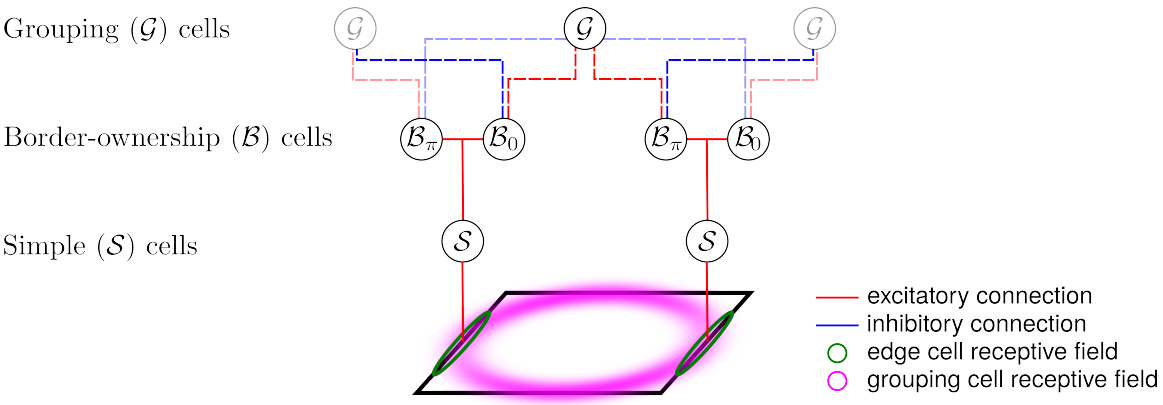


Figure 3.3: Structure of the recurrent neural network. Each circle stands for a population of neurons with similar receptive fields and response properties. Red and blue lines represent excitatory and inhibitory projections, respectively. Solid and dashed lines represent purely feedforward and reciprocal feedforward/feedback connections, respectively. Edges and other local features of a figure (black square outline) activate simple cells ($\mathcal{S}$) whose receptive fields are shown by green ellipses. $\mathcal{S}$ cells project to border-ownership cells ($\mathcal{B}$) that have the same preferred orientation and retinotopic position as the $\mathcal{S}$ cells they receive input from. However, for each location and preferred orientation there are two $\mathcal{B}$ cell populations with opposite side-of-figure preferences, in the example shown $\mathcal{B}_\pi$ whose neurons respond preferentially when the foreground object is to the left of their receptive fields and $\mathcal{B}_0$ whose members prefer the foreground to the right side of their receptive fields. $\mathcal{B}$ cells have reciprocal, feedforward excitatory and feedback modulatory connections with grouping cells, $\mathcal{G}$, which integrate global context information about objects. The receptive field of a $\mathcal{G}$ cell is shown by the purple annulus. Opposing $\mathcal{B}$ cells compete indirectly *via* feedback inhibition from $\mathcal{G}$ cells, which bias their activity and thus generate the border-ownership signal used to determine figure-ground assignment.

as

$$f(x,y) * g(x,y) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f(m,n) g(x+m, y+n) \tag{3.5}$$

v_θ is generated using the von Mises distribution as follows

$$v_\theta(x,y) = \frac{\exp\left[(\sqrt{x^2+y^2} - R_0)\cos(\tan^{-1}(\frac{y}{x}) - \theta + \frac{\pi}{2})\right]}{2\pi I_0(\sqrt{x^2+y^2} - R_0)} \tag{3.6}$$

where R_0 is the radius of the grouping cell receptive field, set to 2 pixels, and θ is the desired angle of the mask. The factor $\frac{\pi}{2}$ rotates the mask to ensure it is correctly aligned with the edge cells. I_0 is the modified Bessel function of the first kind. v_θ is then normalized according to

$$v_\theta(x,y) = \frac{v_\theta(x,y)}{\max(v_\theta(x,y))} \tag{3.7}$$

On the first iteration, since there is no difference in activity between each pair of $\mathcal{B}$ cells as they receive the same initial bottom-up input, we omit the inhibition by the non-preferred $\mathcal{B}$ cells and only use the activity of the preferred $\mathcal{B}$ cells (eqs. 3.3 and 3.4). On subsequent iterations, we include the inhibitory term. We also implement a simple form of local inhibition between the two grouping pyramids, $\mathcal{G}_L^k(x,y)$ and $\mathcal{G}_D^k(x,y)$. At each spatial location, only one type of $\mathcal{G}$ cell should be active, representing either a light or dark object at that location. For each level of the pyramid k, we perform a max operation as follows

$$\mathcal{G}_L^k(x,y) = \begin{cases} \mathcal{G}_L^k(x,y) & if\ \mathcal{G}_L^k(x,y) > \mathcal{G}_D^k(x,y) \\ 0 & otherwise \end{cases} \tag{3.8}$$

$$\mathcal{G}_D^k(x,y) = \begin{cases} \mathcal{G}_D^k(x,y) & if\ \mathcal{G}_D^k(x,y) > \mathcal{G}_L^k(x,y) \\ 0 & otherwise \end{cases} \tag{3.9}$$

Feedback from $\mathcal{G}$ cells to $\mathcal{B}$ cells is used to bias the responses of the $\mathcal{B}$ cells to correctly signal figure-ground assignment. The feedback depends on the contrast polarity of the $\mathcal{G}$ cell and the $\mathcal{B}$ cell.

$\mathcal{B}^k_{\theta,L}$, the border-ownership activity for a light object on a dark background is given by

$$\mathcal{B}^k_{\theta,L}(x,y) = 2\mathcal{S}^k_{\theta,L}(x,y) \times \frac{1}{1+\exp\Big(-\big(\sum_{j\geq k}\frac{1}{2^{j-k}}v_{\theta+\pi}(x,y)*\mathcal{G}^j_L(x,y) - \sum_{j\geq k}\frac{1}{2^{j-k}}v_{\theta}(x,y)*\mathcal{G}^j_D(x,y)\big)\Big)} \tag{3.10}$$

and $\mathcal{B}^k_{\theta,D}$, the border-ownership activity for a dark object on a light background is given by

$$\mathcal{B}^k_{\theta,D}(x,y) = 2\mathcal{S}^k_{\theta,D}(x,y) \times \frac{1}{1+\exp\Big(-\big(\sum_{j\geq k}\frac{1}{2^{j-k}}v_{\theta+\pi}(x,y)*\mathcal{G}^j_D(x,y) - \sum_{j\geq k}\frac{1}{2^{j-k}}v_{\theta}(x,y)*\mathcal{G}^j_L(x,y)\big)\Big)} \tag{3.11}$$

where v_θ is the kernel responsible for mapping object activity in the grouping pyramids back to the objects edges, and the term 2^{-j} normalizes the v_θ operator across scales. The logistic function in the equations above enforces a form of competition between $\mathcal{B}$ cells such that their total activity is always conserved, and each $\mathcal{B}$ cell has activity between the range of zero and two times their initial bottom-up input

activity, $\mathcal{S}^k_\theta(x, y)$.

In the equations above, feedback from $\mathcal{G}$ cells to $\mathcal{B}$ cells is contrast-sensitive. $\mathcal{B}$ cells receive excitatory feedback from $\mathcal{G}$ cells of the same contrast polarity on their preferred side and inhibitory feedback from $\mathcal{G}$ cells of the opposite contrast polarity on their non-preferred side. This is motivated by neurophysiological results which show that image fragments placed within the extra-classical receptive field of a border-ownership neuron can cause enhancement of the neuron's activity when placed on its preferred side, and suppression if placed on the non-preferred side (Zhang and von der Heydt, 2010). Furthermore, modulating the bottom-up $\mathcal{S}$ cell responses with $\mathcal{G}$ cell activity summed across spatial scales ensures that the $\mathcal{B}$ cell responses are scale-invariant. Neurophysiological results show border-ownership coding for stimuli of varying sizes, with the latency of the border-ownership signal being relatively independent of the size of the figure (Zhou et al., 2000; Sugihara et al., 2011).

Figure-ground assignment should be robust for both light objects on dark backgrounds and dark objects on light backgrounds. In our model, this is achieved by computing $\mathcal{B}$ cell activity independently for each contrast polarity and then summing this activity to give a final border-ownership response independent of figure-ground contrast polarity. The $\mathcal{B}$ cell responses for light and dark objects are combined to give a contrast polarity invariant response

$$\mathcal{B}^k_\theta(x, y) = \mathcal{B}^k_{\theta,L}(x, y) + \mathcal{B}^k_{\theta,D}(x, y) \tag{3.12}$$

The sign of the difference $\mathcal{B}_\theta(x,y) - \mathcal{B}_{\theta+\pi}(x,y)$ determines the direction of border ownership at pixel (x,y) and orientation θ. Its magnitude gives a confidence measure for the strength of border ownership.

Similarly, the $\mathcal{G}$ cell responses for light and dark objects are combined to give a contrast polarity invariant response

$$\mathcal{G}^k(x,y) = \mathcal{G}^k_L(x,y) + \mathcal{G}^k_D(x,y) \tag{3.13}$$

The $\mathcal{G}$ pyramid from eq. 3.3 and eq. 3.4, summed across both the light and dark channels, along with the figure-ground assignments in the $\mathcal{B}$ cells, is the output of the grouping algorithm and provides a perceptual organization of the visual scene.

As mentioned previously, we use both luminance and color information from the image in order to perform the grouping operation. The same exact operations that were performed on the luminance channel are also performed on the two color channels. We combine the final outputs of the $\mathcal{B}$ and $\mathcal{G}$ cells with an 80% weighting for the luminance channel and a 10% weighting for both the red-green and blue-yellow color channels. This choice of weighting does not qualitatively change our results.

3.2.2 Model Implementation

All simulations were performed on a 300-core CPU cluster running Rocks 6.2 (Sidewinder), a Linux distribution intended for high-performance computing. This

allowed us to independently run our model on different images, speeding up our testing time. We ran the model for a total of 10 iterations, with each iteration being one feedforward pass of $\mathcal{B}$ cell to $\mathcal{G}$ cell activity, followed by one feedback pass of $\mathcal{G}$ cell to $\mathcal{B}$ cell activity (Figure 3.2). We generally found that the model converged after a few iterations. The final $\mathcal{G}$ cell activity was summed across scales. Contour detection and figure-ground assignment results are computed from the population of $\mathcal{B}$ cells at the highest resolution level of the image pyramid, which had the same resolution as the input image. $\mathcal{B}$ cell activity is converted into a population vector code by summing the final activity across orientations, where the magnitude of the resulting vector represents the border ownership signal (or edge strength), and the direction of the vector provides a continuous figure-ground orientation label. For a given image, we normalize the border ownership signal at each pixel (x, y) by the maximum border ownership signal across the entire image, such that the border ownership signal is bounded between -1 and +1.

We benchmarked our model on the publicly available Berkeley Segementation Dataset (Martin et al., 2001). We did this in the context of two tasks: contour detection and figure-ground assignment. Each dataset includes 100 to 200 test images, and we report F-scores (harmonic mean of precision and recall) for the contour detection task and mean accuracy (percent of correctly labeled figure-ground edges) for figure-ground assignment task averaged over all test images. We used publicly available benchmarking code to do our analysis and comparisons with other approaches.

3.2.3 Model to Cell Comparison

To compare our model with experimental results, we used a publicly available dataset of border-ownership cell responses recorded during viewing of natural scenes (Williford and von der Heydt, 2017). More details about the stimuli, experimental design, and data analysis can be found in the corresponding paper (Williford and von der Heydt, 2016). The dataset includes border-ownership signals (BOS) for each scene that was viewed by each recorded cell. In order to compare our model with the experimental results, we calculated the model's border ownership signal for the same set of scenes shown to the cells. We quantified model performance by using a combination of the cosine similarity metric, regression goodness of fit, and equivalence testing, which are explained in more detail below.

3.2.3.1 Cosine Similarity

To compare the model with cells from the experiment, we first chose a subset of the cells ($N = 13$) that had highly consistent border-ownership responses (defined as having the same side of border ownership on >80% of their tested scenes). To compare the BOS response of a cell to another cell or to the model on the set of common scenes viewed by both, we used the cosine similarity metric, which is defined as

$$\cos(\theta) = \frac{A \cdot B}{||A||_2||B||_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3.14}$$

where A_i and B_i are components of the vectors A and B respectively.

The cosine similarity is bounded between -1 and +1, with the geometric interpretation that the metric measures the cosine of the angle between two vectors in a high dimensional space. Two vectors which are exactly the same will have a cosine similarity of 1, two vectors that are exactly opposite will have a cosine similarity of -1, and a cosine similarity of 0 indicates two vectors that are orthogonal or decorrelated. In our case, we can treat the responses of the model or a given cell on a set of scenes as a vector in a high-dimensional space (with each axis being the BOS for one scene) and we can then compute the cosine similarity between any two vectors (*e.g.* between one cell and another cell or between the model and a cell). We also explored the use of other similarity metrics, such as the Pearson correlation coefficient (r), but found that these metrics often unfairly reduced the similarity measure due to mean centering of the BOS responses.

3.2.3.2 Equivalence Testing

To test the hypothesis that the model performs similarly to cells from the experiment, we use equivalence testing, which has more traditionally been used in the bioequivalence setting, for example, to compare whether the efficacy of a new drug is

similar to that of an existing drug on the market. In standard hypothesis testing, the null hypothesis is that mean of the two distributions are not statistically different. However, failure to reject the null hypothesis is not sufficient evidence to conclude that the two distributions are similar, as the test may also fail due to not having enough statistical power. In equivalence testing, the null hypothesis is instead that the means of the two distributions differ by a pre-determined "zone of scientific significance." The alternative hypothesis (where the burden of proof lies) is that the means of the two distributions are actually the same. In our case, the equivalence test is calculated by either using two one-sided t-tests or computing confidence intervals on the difference between the model-cell and cell-cell cosine similarity distributions, and determining whether this confidence interval lies within the pre-determined zone of scientific significance. We chose a zone between -0.2 - 0.2, which represents $\pm 10\%$ of the full range of possible cosine similarity values.

3.2.3.3 Goodness of Fit

In order to quantify model performance on a per-cell basis, we performed linear regression on the cell's border-ownership responses across scenes (ordinate) against the model's border-ownership responses across the same set of scenes (abscissa). For each regression calculation, we forced the least-squares regression line through the origin because the sign of the border-ownership signal for a given neuron is arbitrary. Two border-ownership selective neurons responding to the same edge can have opposite

side-of-figure preferences, and which direction we assign a positive value is arbitrary. This ambiguity means that the line of best fit must be invariant against reflecting any data points about the origin, which is why the fitted line must pass through the origin.

However, each cell's response contains a repeatable component (in response to the same stimulus and which we attempt to capture with our model) and a noise component (which is completely random). Because our model is deterministic, it is unable to capture the noise component present in the responses of cells. The linear regression provides a measure of the variance that can be explained by our grouping model. However, we only care about the explainable variance, which is the total response variance minus the noise variance, which can be estimated from a small number of stimulus presentations from experiments. As a result, we defined a goodness of fit measure for the regression by computing the fraction of explainable variance that can actually be explained by the model. We follow the general methods put forth in DiCarlo et al. (1998) for calculating the goodness of fit measure.

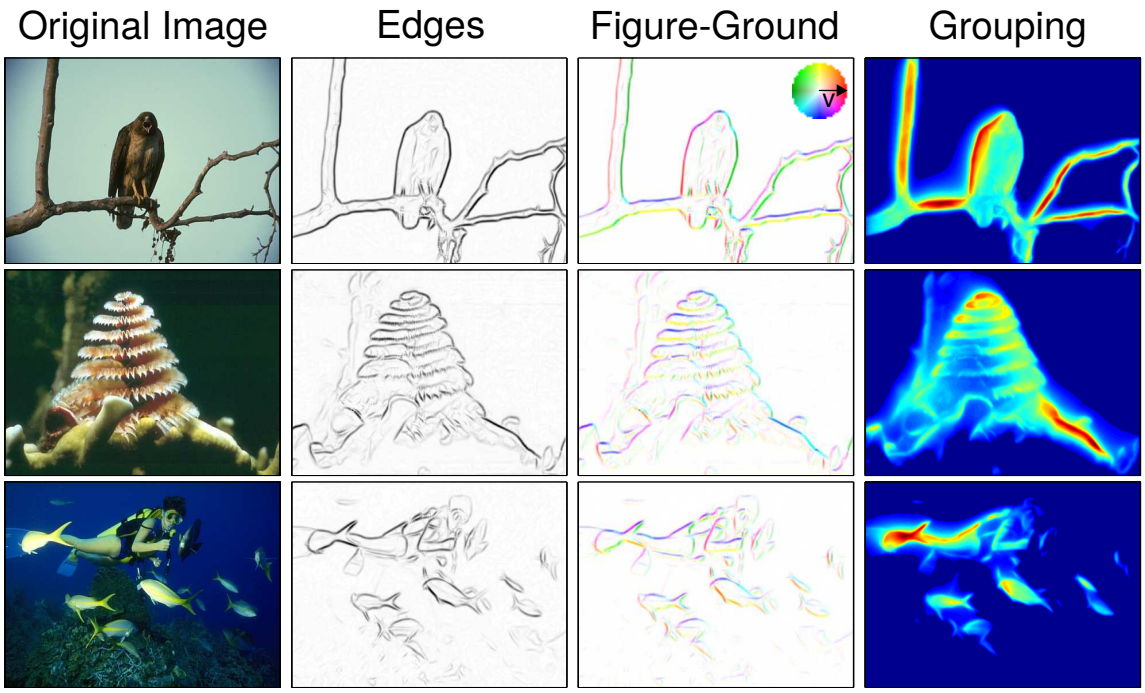


Figure 3.4: Results of our model on example images from the Berkeley Segmentation Dataset. Columns from left to right are the original images, the edge maps, the border-ownership cell activity (representing figure-ground assignment), and also grouping cell activity. For the figure-ground assignment, the color of each edge is represented by a hue and a saturation value (see color wheel insert). The hue of the edge represents the figure-ground orientation label with the arrow convention shown in the color wheel (*e.g.* red represents an object pointing to the right) and the saturation of the edge represents the strength of the border-ownership signal.

3.3 Results

3.3.1 Evaluation of the model on standard benchmarks

We benchmarked our model on the Berkeley Segmentation Dataset (Martin et al., 2001). We did this for two separate tasks: 1) a contour detection task and 2) a figure-

ground assignment task. Importantly, our model uses the same set of parameters for both tasks, and parameters were not tuned separately for each task. Examples of our model output are shown in Figure 3.4. Here, we show the original input image, the edge signals, the border-ownership signals, and the final grouping maps. We hypothesize that the border-ownership signal is a good correlate of the perceptual saliency of object contours. As a result, we use the border-ownership signal (independent of figure-ground orientation) as the output for the contour detection task. We quantify our results on the contour detection task using the F-score, which is the harmonic mean of accuracy and precision of contour detection. For the contour detection task, we compare our approach to three different approaches: ultrametric contour map (Arbelaez et al., 2011, gPb-owt-ucm), structured edges (Dollár and Zitnick, 2015, SE), and structured random forests (Teo et al., 2015, SRF). Overall, we achieved an F-score of 0.64 on the contour-detection task. State-of-the-art computer vision models achieve F-scores on the order of 0.73 (all three approaches gave similar results). Our model does not perform as well as these other approaches, which may be the result of the limitations in the initial edge detection method we used in our model.

For the figure-ground assignment task, we quantify our results using the mean accuracy of figure-ground assignment across all labeled contours in the test images. The model's figure-ground label for a given scene point is considered correct if it falls within 90° of the true figure-ground label. We compared our model to structured random forests (Teo et al., 2015, SRF) and two conditional random field approaches,

	BSDS500		
	ODS	OIS	AP
Human	0.80	0.80	-
Our approach	0.64	0.65	0.51
gPb-owt-ucm	0.73	**0.76**	0.73
SE	0.73	0.75	**0.77**
SRF	0.73	0.74	0.76

Table 3.1: Contour-detection results. Numbers shown are the F-measures when choosing the optimal scale for the entire dataset (ODS) or per image (OIS), as well as the average precision (AP).

	BSDS
	Mean Accuracy
Our approach	71.5%
SRF	**74.7%**
Global-CRF	68.9%
2.1D-CRF	69.1%

Table 3.2: Figure-ground assignment results. Numbers shown are the mean accuracy across all matched scene points.

Global-CRF (Ren et al., 2006) and 2.1D-CRF (Leichter and Lindenbaum, 2009). Overall, we achieved a mean accuracy of 71.5% on the figure-ground assignment task. Structured random forests achieve a mean accuracy of 74.7%. Surprisingly, our model outperforms other approaches based on conditional random fields (Ren et al., 2006; Leichter and Lindenbaum, 2009), which achieved mean accuracies below 70%. There is also a recent deep learning approach to the same problem (Wang and Yuille, 2016), but since the results of this method were not benchmarked using the above standard tests, we do not include it in our comparison. Notably, these computer vision approaches achieve higher performance than our model, but require extensive training and parameter tuning on a held-out training set of images. Although our

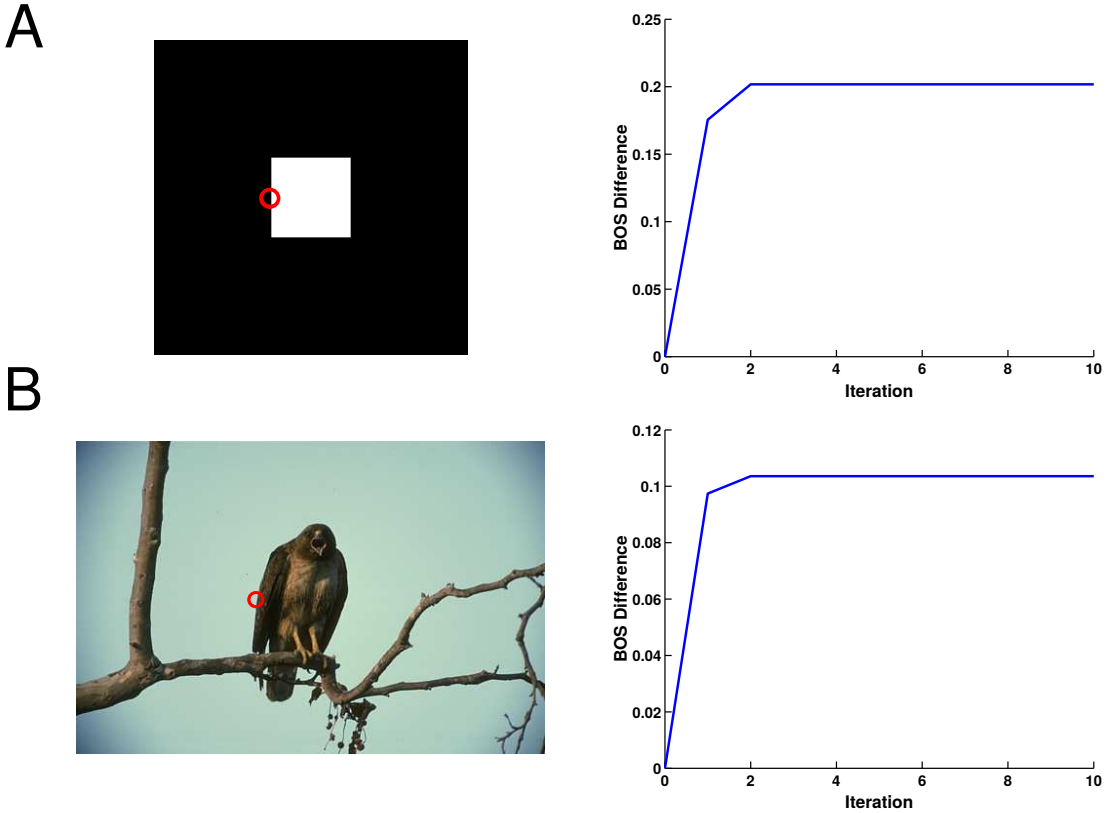


Figure 3.5: Time course of border-ownership coding. The model converges to the correct border-ownership assignment within two-three iterations (one iteration corresponds to a feedforward and feedback pass through the model). The receptive field of the model's border ownership cell is shown by the red circle. The input image and time course of the border-ownership signal are shown for both the standard square commonly used in experiments (A) and an example natural scene from the Berkeley Segmentation Dataset (B). The border-ownership signal is computed as the difference in activity between the preferred and non-preferred border-ownership cells.

model does not outperform state-of-the-art, it does represent an alternative approach based on biologically plausible computations that require very little training or tuning of parameters.

We report our results on the contour detection and figure-ground assignment tasks in Tables 1 and 2 above, respectively. All benchmarks were performed using public

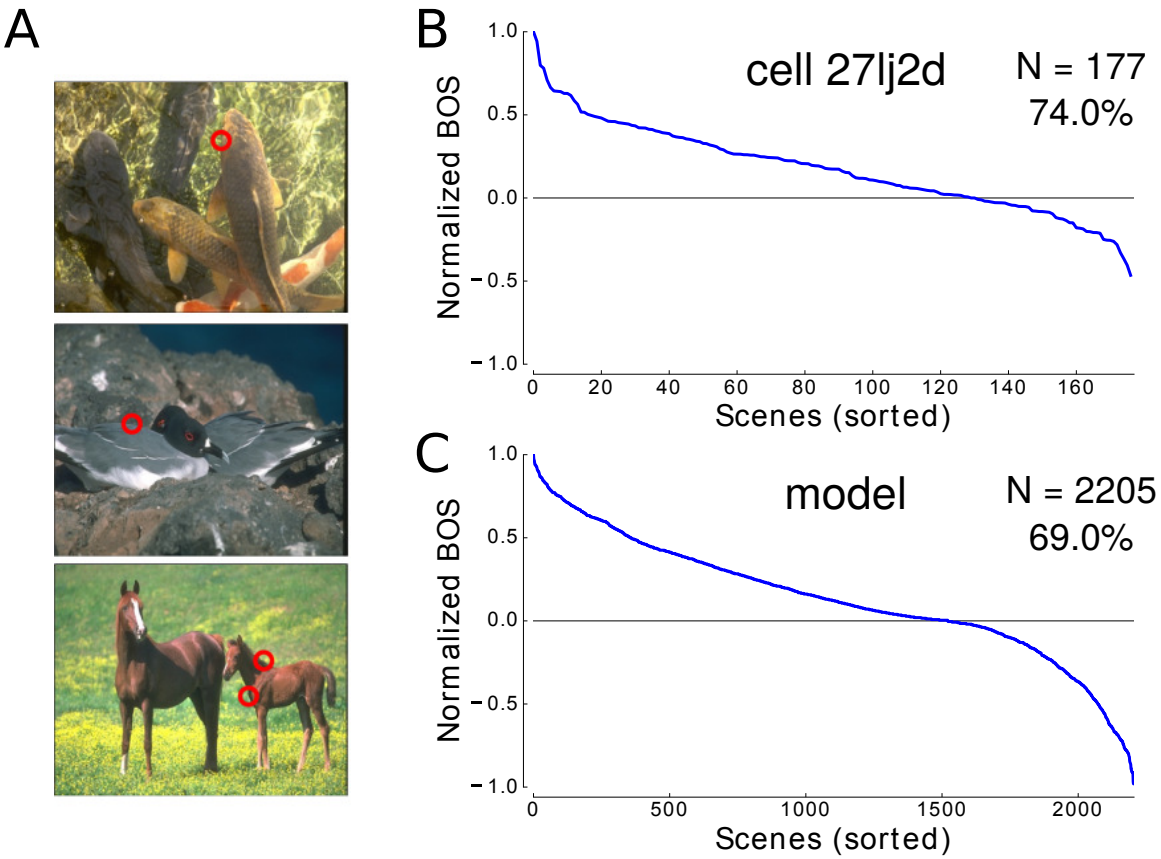


Figure 3.6: Cell and model consistency across scenes. (A) Examples of scenes that were used to test border-ownership selectivity during the experiments. Red circles represent scene points within the images, which were centered on the receptive fields of border-ownership selective neurons during the experiments and during our testing of the model. A single image could contain multiple scene points, as shown by the example in the last row. (B) The normalized border-ownership signal (BOS) for example cell $27lj2d$ is shown according to each scene, with scenes sorted in decreasing order by strength of BOS. The cell achieved a consistency of 74.0% across all tested scene points ($N = 177$). Consistency is defined as the number of scenes with positive border-ownership signal divided by the total number of scenes. (C) The normalized BOS signal for the model is shown with the same convention as in (B). The model achieved an overall consistency of 69.0% across all tested scene points ($N = 2205$), consistent with our finding that the overall accuracy of the model is around 70% on the figure-ground assignment benchmark.

code made available by the authors of the original papers.

3.3.2 Timing of the border-ownership signal

We tested our model on both the standard square stimuli which are often used to determine border-ownership preference in experiments (Zhou et al., 2000), as well as a wide array of natural scenes from the Berkeley Segmentation Dataset. We found that our model converges within a few iterations, demonstrating that only a few feedforward and feedback passes are needed to determine figure-ground assignment for a given image. Given that white-matter projections in the brain are quite fast, we assume that a single feedforward and feedback pass in our model takes about 10 ms. As the model converges within 2-3 iterations, the border-ownership signal will reach its peak within 20-30 ms of the initial visual response. A similar time course has also been observed in the experimental data, with the border-ownership signal appearing approximately 30 ms after visual response onset (Williford and von der Heydt, 2016). The similar time course of BOS tuning on both artificial and natural stimuli points to a common cortical mechanism for grouping, which is also supported by previous experimental results demonstrating consistent border-ownership coding across different stimuli (Figure 3.5).

3.3.3 Comparison of model results to experimental results

The model exhibits consistent border-ownership coding across a large number of natural scenes. Figure 3.6 compares the border-ownership signals sorted in descending order by scene for an example cell with that for the model. The example cell shows a consistency of 74.0% across 177 scenes, which was the largest number of scenes that was tested for any single cell in the dataset. A large number of cells in the dataset were highly consistent, including the 13 cells we chose with >80% consistency, and also 3 cells with >90% consistency. In comparison, the model shows an overall consistency of 69.0% across 2205 scenes, which is more than an order of magnitude more scnes than was tested on the example cell. This level of consistency is similar to the accuracy the model achieved on the figure-ground assignment benchmark, which was also close to 70%.

We also used the cosine similarity metric to quantify similarity in BOS responses between cells and agreement between cells and the model on a shared set of scenes. Despite the large diversity of cells and their responses, we found that our model was able to largely explain the border-ownership coding of highly consistent cells on natural scenes. Figure 3.7 shows the cosine similarities for cell-model comparisons on a per-cell basis for the subset of highly consistent cells. The model-cell cosine similarities were all positive, ranging from 0.21 - 0.69, with a mean similarity of 0.44.

Given biological noise and inter-cell differences, we did not expect that the model-cell cosine similarities would reach 1. In order to characterize an upper bound on the cosine similarity values, we calculated the cosine similarities between all pairs of highly consistent cells($N = 58$ pairs). For the cell-cell comparisons, the cosine similarities ranged from 0.14-0.91, with a mean similarity of 0.54. Equivalence testing on the means of the cell-cell cosine similarities and model-cell cosine similarities revealed no differences in the mean values based on a zone of scientific significance of -0.2 - 0.2 ($p = 0.02$). As a result, we conclude that our model performs similarly to the highly consistent cells in the dataset.

We also computed regression fits between the cell border-ownership responses and model border-ownership responses on a per-cell basis. Each regression outputs an R^2 value, which gives a measure of the percent of variance that the model is able to capture. The noise variance was estimated from the responses of each cell to a given scene and summed over the total number of scenes presented. The R^2 goodness of fit values ranged from 0.07-0.67. For two of the highly consistent cells, the R^2 measure exceed 0.5, indicating that the model was able to capturing over 50% of the explainable variance.

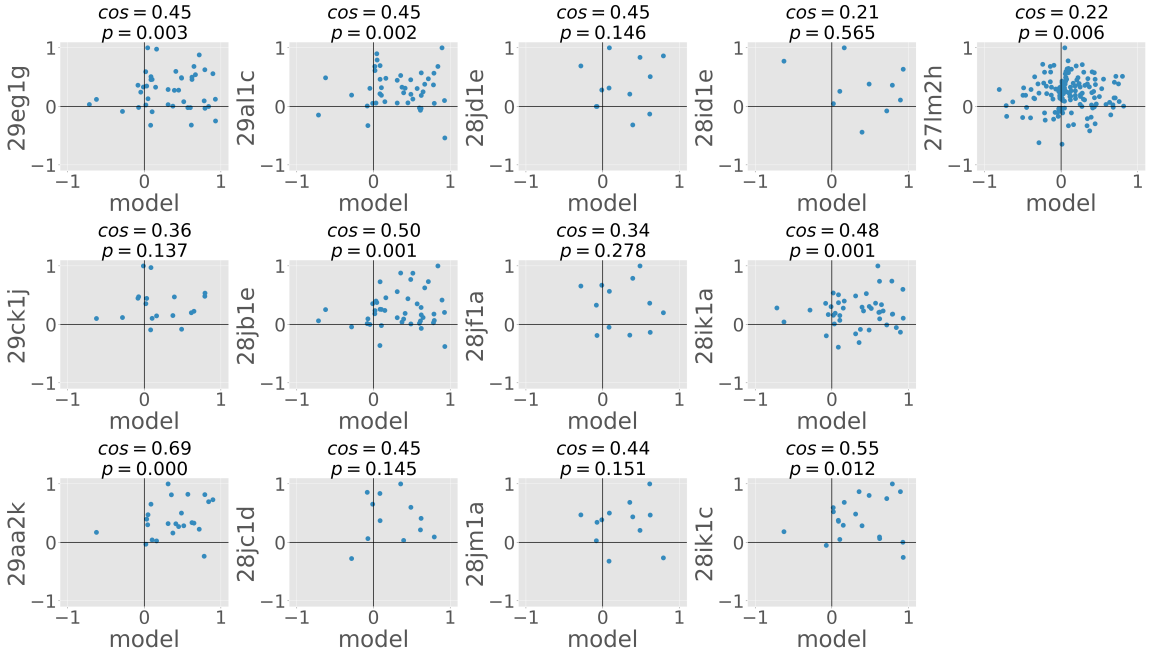


Figure 3.7: Cosine similarity metric comparing the model to the population of consistent cells. Each subplot shows a scatter plot of one cell's border-ownership signal against the model's border-ownership signal on the common set of scenes viewed by both. The cosine similarity metric along with the associated p-values (statistically different from a cosine similarity of zero) are shown above each scatter plot. Cosine similarities for the cell-model comparisons ranged from 0.21 - 0.69, with 7/13 cells having significant cosine similarities.

3.4 Discussion

3.4.1 Understanding the cortical mechanisms of figure-ground organization

We propose that a simple grouping mechanism can explain figure-ground organization in natural scenes. Grouping cells in our model enforce the Gestalt principles of continuity and proximity with their annular receptive fields. Importantly, the design

of these receptive fields was based on first principles, and not due to any training or parameter tuning on natural scenes, as is common in machine learning approaches. We show that this receptive field structure is useful for assigning figure-ground relationships in both artificial and natural stimuli. These receptive fields capture the convex shape of objects, which has been shown to be an important cue from the analysis of natural scene statistics (Sigman et al., 2001). Our model does not rely on local cues, such as T-junctions and X-junctions, or higher-level object identity information, such as that from higher visual areas which may influence segmentation based on object familiarity. Instead, we propose that grouping in our model operates at intermediate levels of the visual hierarchy to structure the visual scene into proto-objects useful for further visual processing.

Our model border-ownership responses match the border-ownership responses of highly consistent cells from the experiments, which was tested by presenting the model with the same set of natural scenes as viewed by the cells. This is surprising given the diversity of cell responses to different natural scenes– cells themselves are not entirely consistent with each other, perhaps indicating that a population of neurons is needed to accurately encode figure-ground relationships (Hesse and Tsao, 2016). However, our model, which is based on the simple principle of annular grouping cell receptive fields, is able to capture the responses of many of these neurons. Furthermore, our model shows high accuracy across all tested scene points in a standard image benchmark dataset. We achieve an accuracy that exceeds 70%, which is close to human

performance on local boundary shapes derived from the same image dataset that we used (Fowlkes et al., 2007). Humans achieve an accuracy of 68-69% on this task. This suggests that the additional feedback connections in our recurrent model, which capture global context information about objects, may improve performance and may be helpful for figure-ground organization.

Our model relies on feedforward and feedback connections via fast white-matter projections between visual areas. This is consistent with the fast appearance of border-ownership signals after visual stimulus onset. This is a clear difference between our model and others which rely either on feedforward or lateral connections. We also use a variety of grouping cells of different scales, which allows our model to achieve relative scale invariance across the range of object sizes present in natural scenes. The main contribution of our present work is the development of a fully-image computable computational model of figure-ground organization that can be applied to natural scenes, which allows for further study of the potential cortical mechanisms of this process.

3.4.2 Comparison to other models

Many have argued that figure-ground assignment is purely local phenomenon that only requires lateral connections (Grossberg, 1994, 1997; Zhaoping, 2005). However, these models are largely untested on natural stimuli, and it remains to be seen if previous results on artificial stimuli will generalize to more difficult real-world con-

ditions. Our model is a member of a broad class of theoretical models that achieve image understanding through bottom-up and top-down recurrent processing (Ullman, 1984; Hochstein and Ahissar, 2002; Roelfsema, 2006; Epshtein et al., 2008). Our model is explicit in that feedback connections from higher visual areas modulate the responses of early feature-selective neurons involved in the related processes of contour integration and figure-ground segregation. Despite requiring feedforward and feedback passes of information through the model, our model converges quite quickly, consistent with the fast establishment of figure-ground relationships in visual cortex.

The only other model that has been tested on natural stimuli involved feedforward processing of asymmetric surround contrast (Nishimura and Sakai, 2005; Sakai et al., 2012). In contrast to our model, their approach was not completely bottom-up (and hence, not fully image-computable). Instead, Sakai et al. (2012) tested model performance on human-labeled contours from the Berkeley Segmentation Dataset. Furthermore, their model was only able to use luminance information, so all images were first converted to grayscale. Our model is fully image-computable, which means that it can be applied to any image, even those without human-labeled contours. Our model is also able to incorporate both luminance and color information from images, which allows it to be applied on a larger range of images and also to study the relative influence of these two cues for grouping.

Importantly, the two models also differ in their predictions about the role of feedback in figure-ground segmentation. One experimental prediction of our model

is that disrupting feedback from higher visual areas (specifically, the feedback from grouping cells) would impair the figure-ground assignment process, and potentially result in poor border-ownership assignment and segmentation of objects in the scene. Models based purely on feedforward processing do not make this prediction. We also predict the existence of contrast-senstive and color-sensitive grouping cells, which send reciprocal feedback connections to similarly-tuned border-ownership cells.

3.4.3 Grouping neurons

There is no clear neurophysiological evidence for grouping neurons yet, although previous studies have found neurons in V4 that respond to contour segments of various curvatures (Pasupathy and Connor, 2002; Brincat and Connor, 2004). The receptive fields of these neurons are similar to those proposed in the model by Craft et al. (2007). Other types of grouping neurons may also exist, including those that respond to gratings (Hegdé and Van Essen, 2007), illusory surfaces (Cox et al., 2013), or 3D surfaces (He and Nakayama, 1995; Hu et al., 2015). We do not attempt to model the whole array of grouping neurons that may exist, but only those necessary for reproducing figure-ground assignment in natural scenes. Furthermore, grouping neurons may be cross-modal, in that they respond to different features that may aid the scene segmentation process, such as disparity, motion, *etc.*. In fact, experimental results show that border-ownership selective neurons have consistent border-ownership tuning across 2D luminance and 3D disparity cues (Qiu et al., 2005). We have not yet

incorporated these additional features into our model, but this represents a potential future area of research.

3.4.4 Scope and limitations of the model

Our model assigns distinct roles to the different visual areas, *e.g.* edge processing in V1 by simple cells, figure-ground assignment in V2 by border-ownership selective cells, and grouping of objects in V4. However, the physiological properties of neurons in early visual areas have not been fully characterized, and neurons in these different areas may have additional ranges of selectivity than the ones we assign them in our model. Our model also produces a rough approximation of the time course of border-ownership coding through a discrete iterative process. As such, it does not allow us to study the dynamics of the recurrent network at a finer time scale. For example, the synchrony between border-ownership neurons that are part of the same object is of particular interest (Martin and von der Heydt, 2015; Wagatsuma et al., 2016). Furthermore, we focused more closely on the border-ownership cell activity in our model and did not specifically study the grouping cell responses of our model, but the combined activity of grouping cells across cells can be used to study a wide range of other visual phenomena, including object segmentation and visual saliency.

Chapter 4

Neural models for the perceptual organization of 3D surfaces

4.1 Introduction

Since we live in a complex 3D world, competent interaction with the surrounding 3D scene structure is indispensable to us and our machines. Access to information about surfaces present in the scene allows us to perform a wide variety of tasks, ranging from motor planning (*e.g.* reaching for a cup a certain distance away on the table) to spatial navigation (*e.g.* following directions in a new city environment).

In studying perceptual organization, researchers have traditionally relied on simple 2D stimuli such as oriented bars (Palmer, 2002). Results from these studies provide support for the importance of well-known Gestalt principles (Koffka, 1935;

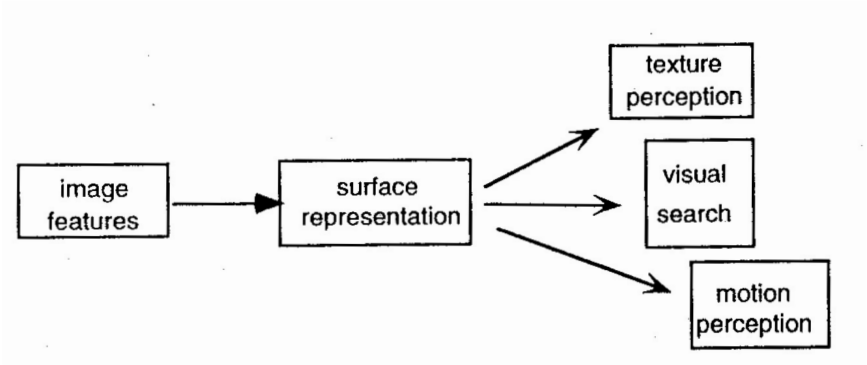


Figure 4.1: The proposed view in which surface representation precedes perceptual functions such as texture perception, visual search, and motion perception. Reproduced from Nakayama et al. (1995).

Wertheimer, 1923), for instance that visual elements are grouped together in a way that begins to give meaning to the visual scene (*e.g.* figure *vs.* background). However, it is unclear how well the results from these relatively simple experiments generalize to the 3D objects and scenes regularly encountered in natural settings. Because we act in a 3D world, perceptual organization must also help to arrange 3D information in a way that can guide our actions.

Perceptual organization also provides a structure for selectively attending to groups of objects (Treisman and Gelade, 1980). Supported by extensive psychophysical data, Nakayama et al. (1995) proposed that surface representations play a key role in intermediate-level vision, providing a critical link between lower-level and higher-level visual functions. In the proposed view, image features give rise to surface representations that can be used for a variety of tasks, including texture perception, visual search, and motion perception (Figure 4.1). For example, subjects can perform efficient search for a conjunction target by selectively attending to the surface that the

target is on in 3D space (Nakayama and Silverman, 1986). Attention has also been shown to spread automatically across surfaces in a separate cueing experiment (He and Nakayama, 1995). These abilities indicate powerful mechanisms for grouping objects into surfaces in 3D space, and suggest that structuring the world in terms of surfaces might be an ecologically important function. These results also have implications for the internal representation of surfaces, because they imply that the visual scene is processed in a way that preserves its 3D structure. This representation must also be able to bring together information from different sensory modalities (*e.g.* vision, audition, *etc.*), in order to form a common representation of the 3D environment that is useful for an agent's behavioral goals (Lewicki et al., 2014).

In addition to being studied by human psychophysical approaches, perceptual organization has been the subject of many studies in the visual system of non-human primates. Many neurons in early visual cortex encode the side to which an object border belongs, a phenomenon known as border ownership (Zhou et al., 2000). Selectivity for the side of ownership involves integrating global context information about the object. Several models have been proposed (Zhaoping, 2005; Craft et al., 2007) to describe how a neuron's border ownership selectivity can be modulated by visual input far away from its classical receptive field with the observed high specificity of object details. One view is that this contextual input is provided by feedback connections from "grouping cells" (Craft et al., 2007) which bias the activity of border ownership cells and thus generate their context-dependent responses. Mihalas et al.

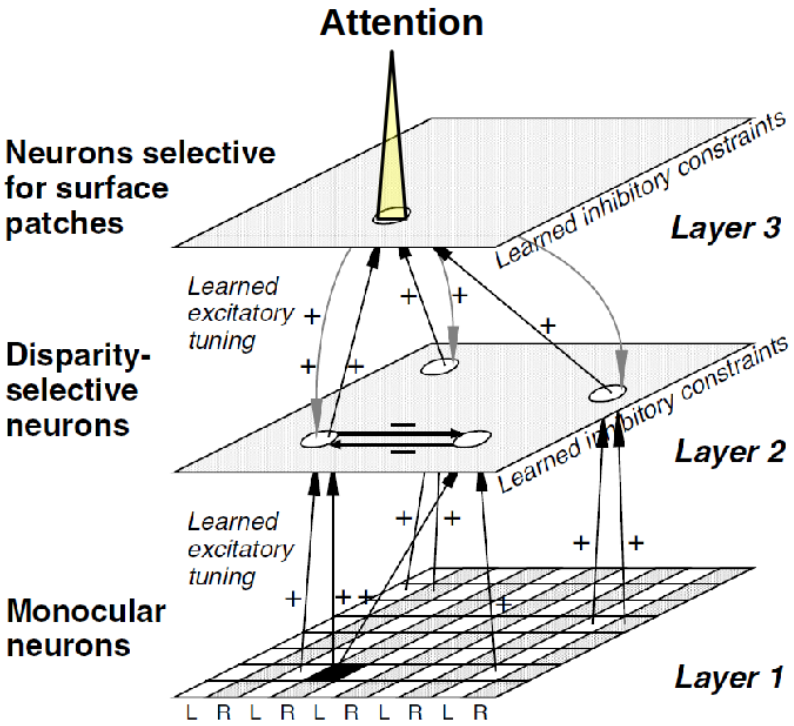


Figure 4.2: Network structure, adapted from Marshall et al. (1996).

(2011) have also shown that grouping cells can direct and sharpen a broad attentional spotlight to the lower-level features of a specific object. In this chapter, we extend this grouping framework to 3D space to show how oriented 3D elements can be grouped into planar surfaces.

Currently, we know very little about how surfaces are represented in the brain, and how this representation is computed. Our model sheds light on a possible neural representation of 3D surfaces and relates this model to previous psychophysical results. Furthermore, we propose that the brain may use basis functions to flexibly and efficiently represent different surfaces using the same population of neurons.

4.2 Methods

4.2.1 Surface grouping model

An overview of the network structure of our model is shown in Figure 4.2. We extend a neural model of visual stereomatching (Marshall et al., 1996) that is conceptually similar to grouping models previously proposed for 2D stimuli (Craft et al., 2007; Mihalas et al., 2011; Russell et al., 2014). The model contains three layers of neurons. The first layer consists of monocular cells, which respond to visual features (e.g. spots, edges, *etc.*) presented to either the left or right eye. The input to the model consists of pairs of stereo images, as would be seen by the left and right eyes. In our model, we set the image input to a value of unity wherever a stimulus is present and zero elsewhere.

The second layer consists of binocular cells, which receive excitatory input from monocular cells. These cells are tuned to a certain disparity based on a fixed spatial weighting between the left and right monocular cells, analogous to the disparity-selective binocular neurons in visual cortex of monkeys (Poggio and Fischer, 1977; Poggio and Poggio, 1984) and cats (Bishop and Pettigrew, 1986; Ohzawa et al., 1990). Lateral inhibition between cells representing different disparities along the same left- or right-eye line of sight reduces potential false matches (Marr and Poggio, 1976).

The third and final layer consists of planar grouping cells which receive excitatory input from populations of disparity-selective cells. Receptive fields of the planar

grouping cells are relatively broad and non-specific, resembling surface patches with a certain range of depth and orientation selectivity in 3D space. These cells may correspond to neurons found in parietal cortex, which have been shown to be selective for the tilt and slant of planar surfaces (Rosenberg et al., 2013). In our model, we used a total of 15 planar grouping cells (*i.e.* five frontoparallel planes, five back slant planes, and five front slant planes), which was sufficient to "tile" the whole 3D visual scene. Planar grouping cells compete with each other through lateral inhibition, which helps to select the best possible interpretation of surfaces within the scene. Additionally, planar grouping cells send reciprocal feedback connections to the disparity-selective cells that define their surface, akin to the relationship between grouping cells and border ownership cells in models of 2D scenes (Craft et al., 2007; Mihalas et al., 2011). To avoid uncontrolled feedback excitation, feedback is multiplicative and only amplifies existing feedforward excitation. Selective attention is modeled as an additive input to those planar grouping neurons representing attended objects. This attentional modulation input is set to a value of 0.25 of the sensory input.

4.2.2 Model equations

All model neurons are simulated as single compartment units with an activity that is modeled as a continuous variable (rate coding). These units are zero-threshold, linear neurons which receive excitatory and inhibitory current inputs. The activity of the units is determined by a set of coupled, first-order nonlinear ordinary differential

equations, which can be solved in MATLAB (MathWorks) using standard numerical integration methods (Euler, Runge-Kutta, *etc.*)

The dynamics of each layer 2 and layer 3 neuron can be described by,

$$\tau f'(t) = -f + [\Sigma W]_+ \tag{4.1}$$

where f represents the neuron's activity and τ its time constant ($= 10^{-2}$ s), W is the neuron's inputs, and $[\ ]_+$ means half-wave rectification.

The feedforward weights from a layer 1 to layer 2 neuron, W_{12}, is

$$W_{12} = \exp(-(\frac{(X_1 \pm D - X_2)^2}{\sigma_{12,x}^2} + \frac{(Y_1 - Y_2)^2}{\sigma_{12,y}^2})) \tag{4.2}$$

where D is the disparity of the layer 2 neuron.

The feedforward weights from a layer 2 to layer 3 neuron, W_{23}, is

$$e_x = cos(S)cos(T)(X_2 - X_3) + cos(S)sin(T)(Y_2 - Y_3) - sin(S)(D_2 - D_3) \tag{4.3}$$

$$e_y = -sin(T)(X_2 - X_3) + cos(T)(Y_2 - Y_3) \tag{4.4}$$

$$e_z = sin(S)cos(T)(X_2 - X_3) + sin(S)sin(T)(Y_2 - Y_3) + cos(S)(D_2 - D_3) \tag{4.5}$$

$$W_{23} = \exp(-(\frac{e_x^2}{\sigma_{23,x}^2} + \frac{e_y^2}{\sigma_{23,y}^2} + \frac{e_z^2}{\sigma_{23,z}^2})) \tag{4.6}$$

where S and T are the slant and tilt, respectively, of the layer 3 planar grouping

neuron.

The lateral inhibition weights between two neurons in either layer 2 or layer 3, W_{ij}, is proportional to the amount of overlap in their feedforward inputs,

$$W_{ij} = \sum_{k \epsilon layer(l-1)} min(W_{ki}, W_{kj}) \tag{4.7}$$

The feedback weights from a layer 3 to layer 2 neuron, W_{32}, are the reciprocal of the feedforward weights, W_{23}, and the feedback is multiplicative in that it only modulates the feedforward input,

$$W_{32} = W'_{23} \tag{4.8}$$

$$E = B(1 + T) \tag{4.9}$$

where E is the net excitation, B is the bottom-up feedforward excitation, and T is the top-down feedback excitation for a layer 2 neuron.

A summary of the critical parameter values for the surface grouping model is shown below in Table 4.1.

4.2.3 Basis function units

Orientation selective cells in V2 have gain field-like properties, with their responses to stimulus orientation depending nonlinearly on disparity (von der Heydt et al.,

Table 4.1: Parameter Values

Parameter	Value	Description
X	1 to 35	x-position (range)
Y	1 to 35	y-position (range)
D	-7 to 7	disparity (range)
$\sigma_{12,x}$	0.69	spread in x (layer 1 to 2)
$\sigma_{12,y}$	2	spread in y (layer 1 to 2)
S	0, $\pi/4$, $\pi/2$	slant of grouping neuron
T	$\pi/2$, $3\pi/2$	tilt of grouping neuron
$\sigma_{23,x}$	0.095	spread in x (layer 2 to 3)
$\sigma_{23,y}$	0.045	spread in y (layer 2 to 3)
$\sigma_{23,z}$	1.185	spread in z (layer 2 to 3)

2000). Pouget and Sejnowski (1997) showed that neurons in parietal cortex with gain fields can be used as basis functions of their sensory inputs for mapping different eye-hand coordinate transformations. Using identical mathematics, we propose that visual disparity (or depth) can play the same role as eye position, generating representations of planes of arbitrary orientations in 3D space. If a surface, S, is a nonlinear function of binocular disparity, $\vec{d}$, and retinotopic position, $\vec{r}$, it can be generated by a linear combination of basis functions spanning these two dimensions,

$$\mathcal{S} = \sum_i c_i B_i(\vec{d}, \vec{r}) \tag{4.10}$$

where c_i are cofficients that can be learned depending on the surface, S, and B_i are a set of basis function units with biologically-plausible receptive fields.

Figure 4.3 shows how the space of retinal position and depth is spanned by disparity-selective cells, which forms a set of basis functions. One axis is along binocular disparity ($\vec{d}$), the other is along retinal coordinates ($\vec{r}$); the latter are obviously two

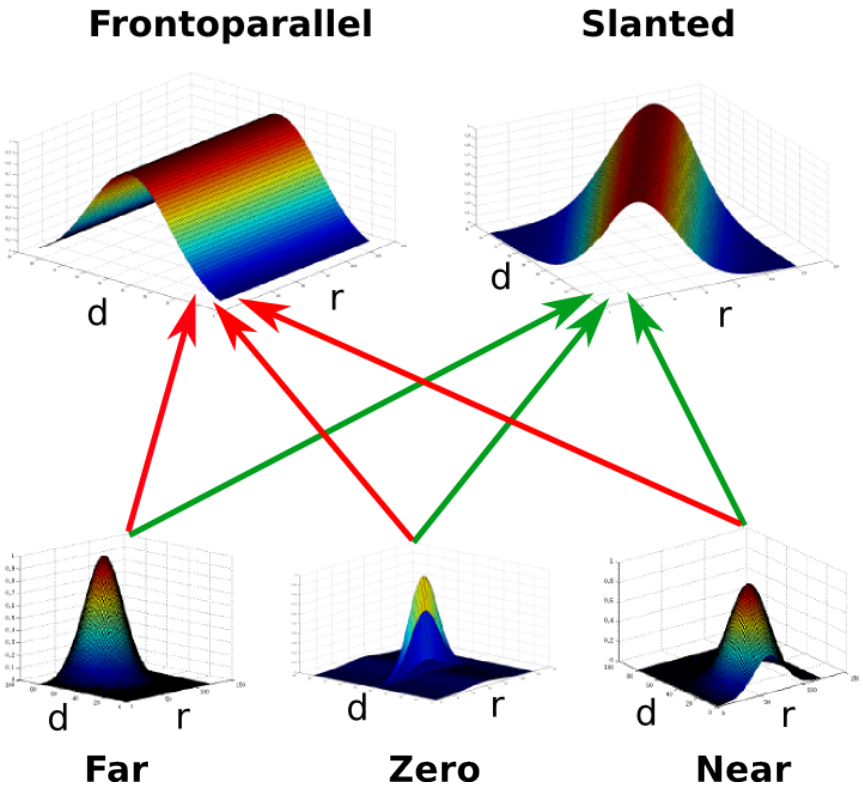


Figure 4.3: The same set of basis functions (*e.g.* far, zero, and near-tuned units with disparity and retinotopic tuning) can be used to generate different surfaces. Learned weights can be used to represent a fronto-parallel surface (red arrows) or a slanted surface (green arrows).

dimensional but only one dimension is shown. Three examples of disparity-selective cells are shown in the bottom row of the figure. From left to right, shown are receptive fields of three foveal cells selective for far disparity, central disparity, and near disparity, respectively. Obviously, these are examples only and a much larger number of cells is needed to span the space in detail. For our results, we used a total of 289 basis function units, B_i , whose receptive fields were computed by multiplying Gaussian retinal receptive fields with disparity tuning curves. The peaks of the Gaussians were spread uniformly between $-4°$ and $4°$, in increments of $0.5°$. The standard deviation

of the Gaussian was fixed at 0.6°. The peaks of the disparity tuning curves were based on the minimal number of neurons required to reproduce psychophysical disparity discrimination curves, which was found to be 17 (Lehky and Sejnowski, 1990). The exact formulation for the near, tuned, and far disparity tuning curves can be found in Table 1 of Lehky and Sejnowski (1990).

Arrows in the figure indicate weighted connections that combine activity of basis function sets into new receptive fields of which two examples are shown at the top. The RF on the left is in the fronto-parallel plane at zero disparity. Planes with non-zero disparities (not shown) are parallel to it but displaced to the left or right in the figure. Top right is a RF that represents a surface slanted in space. In our model, neurons with receptive fields as in Figure 4.3 group features on surfaces in 3D space, in complete analogy to the grouping mechanisms in 2D employed in previous work (Craft et al., 2007; Mihalas et al., 2011). The weights of these connections, c_i , can be learned by traditional supervised learning techniques, *e.g.* the delta rule (Widrow and Hoff, 1960). The training set was composed of 49 pairs of retinal position and disparity, selected from within the range $-3°$ and $3°$ for both values. Weights were adjusted until the mean squared error (MSE) of our approximation to the actual values was less than 0.001.

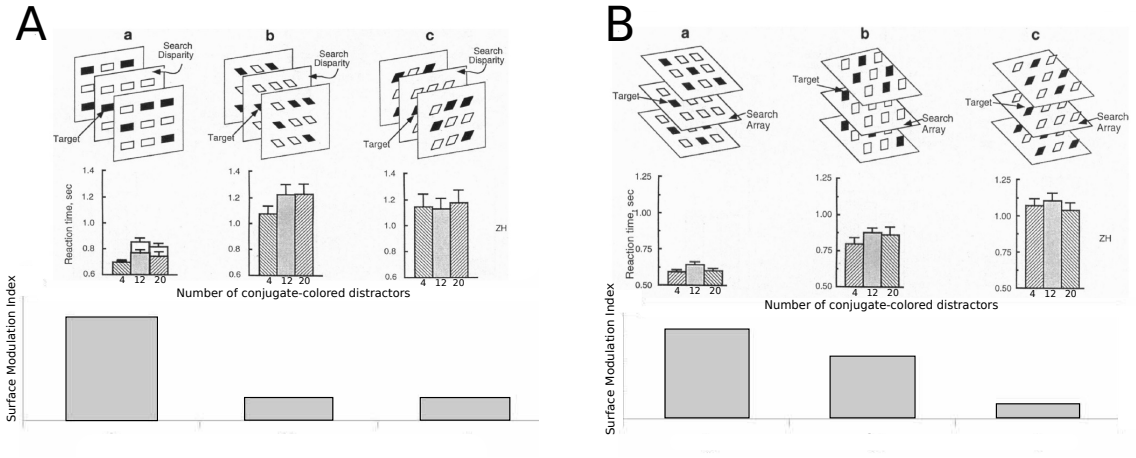
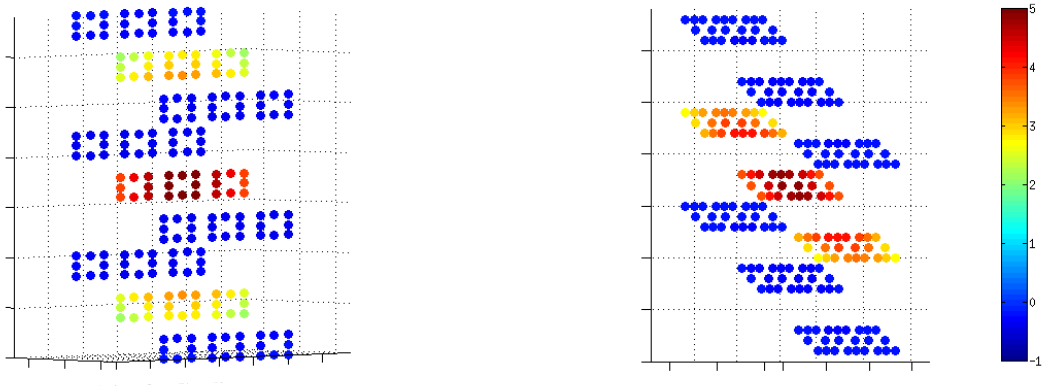


Figure 4.4: Psychophysical and model results, adapted from He and Nakayama (1995). For each trial type (A or B), the top row shows the different stimuli, the middle row shows representative reaction times, and the bottom row shows the degree of attentional modulation of disparity-selective cells on the attended plane. Increase in activity is assumed to be inversely proportion to reaction times.

4.3 Results

4.3.1 Comparison of model to experiment on attention-to-surfaces task

Figure 4.4 illustrates the experimental paradigm of He and Nakayama (1995). The top rows in A and B show schematically the stimulus arrays in which oriented stimulus bars are grouped in different depth planes. In Figure 4.4A, subjects had to search for the odd-colored target in the middle depth plane; planes are outlined by rectangles that were not visible to the observers. The target was unique in this search plane but visually identical distractors were present in other depth planes. In A-a, objects are aligned with the search plane while in A-b and A-c, they are slanted out



(A) (B)

Figure 4.5: Spread of attention across surfaces. (A) When attention is directed to the center fronto-parallel plane (as in the experiment in Figure 4.4A-a), attention enhances the activity of all cells along the surface (red), while suppressing the activity of cells belonging to other surfaces (blue). (B) Similar results also hold for slanted surfaces in the experiment, Figure 4.4B-a. The color bar to the right shows attentional modulation over baseline activity, with reddish colors indicating enhancement and bluish colors indicating suppression.

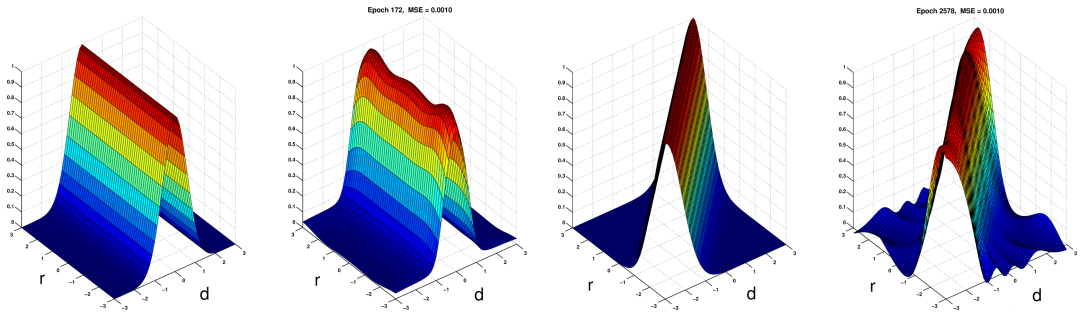
of the plane. The middle row in A shows measured reaction times for three different numbers of distractors.They are significantly shorter when the objects are coplanar with the search plane (A-a) compared to when they are not aligned with the plane (A-b and A-c).

In our model, we assume that the visual system can selectively direct attention towards a specific surface by providing additional excitatory input to the grouping cell that represents this surface. As shown in Figure 4.2, activity from the grouping cell selectively feeds back to all objects on that surface. In the case of Figure 4.4A-a, the grouping cell corresponding to the middle fronto-parallel plane receives this attentional input. Activation of the disparity-selective cells in the search plane, shown

in Figure 4.4A-a, bottom row, is thus high. Among the objects in the search plane, the target has a unique color, which results in efficient search and the target being identified immediately. Reaction times are difficult to simulate in detail, therefore the increase in mean activity of disparity-selective cells on the attended plane due to attentional modulation is plotted instead, which is assumed to be inversely proportional to reaction times. The high activation level, bottom row, translates thus into short reaction times, middle row.

In contrast, in Figure 4.4A-b, the search plane is no longer a well-formed surface but contains objects that are slanted backwards. Directing attention to the middle fronto-parallel plane then has little effect on the disparity-selective cells in the search plane. Search therefore cannot occur entirely within a single plane of coplanar, grouped elements and becomes inefficient, with much higher reaction times (middle row). These long reaction times are reflected in the low activation level of the disparity-selective cells (bottom row). Figure 4.4A-c shows the analog result for figure elements that are slanted forward rather than backwards, as in A-b. Again, reaction times are long and population activity is low.

The result is not restricted to fronto-parallel planes, as similar reaction time results were also found for slanted planes in Figure 4.4B. When subjects are instructed to search in a plane that is coplanar with the orientation of the figure elements (middle plane in B-a), search is fast (B-a, middle row). In contrast, when the figure elements do not align with the search plane (top row in B-b and B-c), reaction time is increased



(A) (B)

Figure 4.6: Using the same set of basis functions, weights can be learned to approximate different surfaces. (A) Fronto-parallel surface (retinotopic location varies while disparity is constant). The actual surface is shown to the left, and the surface approximation is shown to the right. (B) Similar results also hold for the representation of slanted surfaces (retinotopic location varies with disparity). In each panel, only one retinotopic axis is modeled for ease of visualization.

(middle row). An unexpected result was that activation levels decrease with increasing angle between search plane and stimulus bars for search in slanted planes (difference between Fig 4.4B-a and B-c). This may explain the observed differences in reaction time data, although He and Nakayama (1995) did not comment on these differences. Again, under the assumption that reaction times are inversely related to reaction times, the model reproduces human behavior (bottom row).

4.3.2 Spread of attention across surfaces

In our model, attention can be directed to surfaces, and this is implemented as a top-down input to specific surface grouping neurons. With attention, the neurons encoding the targets within a search array had higher activations when the targets were coplanar with each other and formed a congruent surface, which is consistent with

the findings presented in the previous section (He and Nakayama, 1995). For both the fronto-parallel surface (Figure 4.5A) and the slanted surface (Figure 4.5B), top-down attention increased the activity of neurons along the attended surface (shown by the reddish colors), and inhibited the activity of neurons that were not part of the attended surface (shown by the bluish colors). Importantly, the spread of attention was constrained to the surface being attended. In our model, surface grouping neurons are reciprocally connected to disparity-selective neurons coding for targets on a surface. The combination of grouping cells and feature-selective cells represents the presence of "proto-objects" in the scene (Rensink, 2000), which results in a structured perceptual organization of the scene in terms of surfaces.

4.3.3 Surface representation using basis functions

In the previous set of results, we represented surfaces using neurons tuned for every possible retinotopic location and disparity. However, this representation is computationally expensive and does not scale well as the size of the input space increases (*i.e.* the number of neurons grows exponentially). Instead, we propose that basis functions may provide a compact and efficient means to represent surfaces using a limited number of neurons. Using our proposed basis function units, which are a function of retinotopic location and disparity (Section 4.2.3), we were able to learn the weights for two different types of surfaces.

Figure 4.6 shows the surface representations approximated by our basis function

units. We first determined a set of weights that could approximate the fronto-parallel surface (Figure 4.6A). The final approximation had a MSE of 0.001, demonstrating that the units we used form a set of basis functions that could accurately approximate the surface. We then repeated the procedure to determine a set of weights that could approximate the slanted surface (Figure 4.6B), which again reached a mean squared error of 0.001. Our results show that both the fronto-parallel and slanted surfaces could be flexibly represented using the same set of basis functions. In our model, these two surface representations coexist, along with an infinite number of other other potential surfaces, which would be represented by different sets of weights. Importantly, any surface can be represented by a single learned linear projection of the basis function units.

4.4 Discussion

As in our previous models, grouping cells serve as "handles" for selective attention but now attention is directed not to individual objects but to sets of objects organized in planes. Reaction times were fastest when the search array was a well-formed surface defined by locally coplanar elements. When search array elements were slanted away from this surface, reaction times increased. These results suggest that attention is linked to and spreads across perceived surfaces, which organize the visual scene (Figure 4.5). Our model also makes non-trivial predictions about mean firing rates.

Responses of neurons to stimuli in 3D planes should become enhanced, relative to when the stimuli are not part of perceived planes. Furthermore, top-down attention (when cued to a certain plane) provides additional input to G cells and the feedback should enhance responses of neurons in the attended vs. unattended planes.

4.4.1 Extension to other surface grouping phenomena

Grouping implies segregation of some units from others: figure *vs.* ground, vase *vs.* face, or, in this case, objects in one plane *vs.* objects in another plane. In our model, the combination of feedforward grouping with local inhibition allows for local winner-take-all behavior: objects within a plane (or close to it, see below) are grouped together and this group is segregrated from objects further away. If the latter are outside the inhibitory range of the first plane and, of course, arranged themselves in a plane, they will be grouped in a plane of their own. Although we did not explicitly model recurrent local excitation, this may possibly be related to the reverberation needed to explain perceptual persistence. McCarley and He (2001) have found surface priming effects during a similar visual search task, in which the organization of the stimulus array on a previous trial can affect reaction times on the following trial. Related to the idea of local excitation and long-range inhbition, surface grouping cells can excite close grouping cells and inhibit far grouping cells based on

the width of the excitation and inhibition profile. This creates the possibility of depth attraction and repulsion effects, which have been found for isolated dot stimuli as well as random dot stereograms (Stevenson et al., 1991). Extending our model to account for these other phenomena is an area of future research.

4.4.2 Generality of basis functions

Among neurons in striate and extrastriate cortex there are many that respond selectively to stereoscopic disparity (Poggio and Fischer, 1977; Cumming and DeAngelis, 2001) and they must play a role in the organization of stimuli in 3D space. We propose that the exact same theoretical framework proposed for implementing sensorimotor coordinate transformations in parietal cortex (Salinas and Abbott, 1995; Pouget and Sejnowski, 1997) can be used to generate surface grouping cells that represent planes of arbitrary orientations in 3D space. Going from one to the other is a simple affine transformation achieved through combination of neurons with gain field-like receptive fields, and the coefficients of this mapping can be learned by simple, Hebbian-type learning rules. We also note that a surface grouping cell groups all stimuli in the plane it represents, irrespectively of stimulus features (both black and white bars are grouped in planes in Figure 4.4). It remains to be seen whether surface grouping neurons learned using the basis functions we proposed can also model results from the He and Nakayama (1995) experiments.

Even though surfaces can be represented efficiently with basis functions, one might

expect a problem due to the potential inflation of the number of grouping cells to cover all of 3-dimensional space. We have argued that to tile the 2D plane, the number of grouping cells is at least two orders of magnitude smaller than that of B cell (Craft et al., 2007). Even assuming this ratio, the situation seems more difficult in 3D: it appears as if an exceedingly large number of G cells would be required if each plane of any possible orientation (tilt and slant) and any possible depth needs to be explicitly represented by a grouping cell (or a population thereof). This would be a problem for the biological system as well as for our computational model.

Although the number of physical depth planes is obviously infinite, there is no need for them to be explicitly represented at any given time. Indeed, behavioral evidence indicates that only a limited number of depth planes can be perceived in any given situation, up to six with experienced observers being allowed several seconds viewing time, allowing for many eye movements (Tsirlin et al., 2008). The brain must have finer-grained representations of objects in space but those are likely in structures like parietal cortex or hippocampus (Manns and Eichenbaum, 2009; Deshmukh and Knierim, 2013) and not in visual cortex. We are not aware of any quantitative studies that have measured the number of tilted, or slanted, planes that can be perceived simultaneously, but from our model, we predict that this number is small, too. This is an experimental prediction that should be tested.

4.5 Conclusion

Using a simple model of perceptual organization in 3D, we are able to reproduce psychophysical results from a visual search task that required allocation of selective attention to surfaces within the scene. The same grouping cells which organize the scene into planes also act as "handles" for top-down selective attention, enhancing the activity of coplanar elements belonging to the plane. Competition between grouping cells results in surface enhancement of the plane corresponding to the attended grouping cell, and suppression of other planes within the scene. Our proposed surface representation aids visual processing by providing a critical link between low-level visual features and high-level object representations.

We also show that basis functions, which have traditionally been used to model coordinate transformations between different reference frames, can also provide a flexible and efficient representation of surfaces. Using the same set of basis function units tuned for retinotopic location and disparity, we can potentially represent an infinite number of surfaces using a learned set of weights. This provides a powerful theoretical framework for how the brain may encode a large number of possible surfaces with only a small population of neurons.

Chapter 5

3D proto-object based saliency

5.1 Introduction

The brain receives large amounts of visual information that it must make sense of in real-time. Processing the entire visual field with the same level of detail present at the fovea would be an exceedingly complex and costly task requiring much greater computational resources than are available to the brain (Tsotsos, 1990). As a result, primates select only the most relevant information and discard the rest, a process known as selective visual attention. Many models of visual attention are constructed with a bottom-up architecture and rely on local contrast in low-level features such as intensity, color, orientation, or motion. Biologically-plausible center-surround differences across different feature channels of an input image can be used to compute a "saliency map" whose maxima indicate where selective attention is deployed (Koch

and Ullman, 1985; Niebur and Koch, 1996; Itti et al., 1998).

However, there is both psychophysical (Einhäuser et al., 2008) and neurophysiological (Zhou et al., 2000; Qiu et al., 2007) evidence that attention relies not only on these simple image features, but also on the perceptual organization of the visual scene into tentative objects, or proto-objects (Rensink, 2000). Biologically-inspired models of proto-object based saliency have been proposed that take into account these recent findings (Craft et al., 2007; Mihalas et al., 2011; Russell et al., 2014). These models include border-ownership selective cells (referred to as border-ownership cells in the following) and grouping cells, which interact to achieve figure-ground segmentation of the image into proto-objects (figures) and the background (ground).

Border-ownership cells have been found in primate visual cortex, with the majority of neurons in area V2 having this property. These cells signal in their neural activity the one-sided assignment of an object border to the region perceived as figure (Zhou et al., 2000). Border-ownership cells are also modulated by attentional influences (Qiu et al., 2007). Grouping cells integrate global context information about proto-objects in the scene according to Gestalt principles such as closure, continuity, convexity *etc.* Importantly, grouping cells act at intermediate stages of vision and do not require higher-level information about object identity, semantic knowledge *etc.* They send feedback to border-ownership cells *via* fast white matter projections, which bias the activity of border-ownership cells to reflect the correct figure-ground segmentation of proto-objects. In this framework, visual saliency is a function of grouping cell activity,

which represents the size and location of proto-objects within the image.

Border-ownership cells have been shown to respond to figure edges defined by a variety of image features, *e.g.* luminance edges, color edges *etc.* When no monocular edge information is present (*i.e.* when the figures are defined by random dot stereograms using only binocular disparity), border-ownership selectivity is also imparted by stereoscopic edges (Qiu and von der Heydt, 2005). Critically, their response to these different figural cues is typically the same in the two-dimensional (2D) and three-dimensional (3D) cases – the preferred side-of-figure of border-ownership cells is consistent for all cues that define the figure. The activity of border-ownership cells thus provides an interpretation of the visual scene in terms of depth-ordered surfaces that correspond to objects in 3D space. Despite these experimental observations, current models of border ownership do not explicitly use depth information and do not address how traditional 2D Gestalt cues interact with depth cues during the figure-ground segmentation process. An exception is a study by Mishra et al. (2012) who used computer vision methods to compute border ownership from low-level depth information and then performed object segmentation in natural images.

Even though in recent years stereoscopic 3D content has become increasingly prevalent, *e.g.* in the viewing of entertainment programs in cinemas and homes, little is known about how visual attention is deployed within 3D environments. It is thus important to understand how humans allocate their attention when viewing natural images and videos in 3D (Le Callet and Niebur, 2013). Binocular disparity cues, which

can be used to generate strong depth percepts, have been shown to alter different aspects of eye movements when participants viewed 3D images (Jansen et al., 2009) and videos (Huynh-Thu and Schiatti, 2011). Only recently have 3D eye tracking datasets been made available which can be used to compare human eye movements with predictions of attentional models. The availability of these datasets and the recent explosion in new 3D content makes it possible to design computational models of 3D saliency and evaluate their performance objectively.

5.2 Related Work

5.2.1 Models of 3D visual attention

Compared to the number of models that have been proposed for 2D visual saliency, relatively few attempts have been made to study how visual attention is deployed within 3D environments. Existing models of 3D visual attention often compute a 2D saliency map which is then combined with the depth information to produce a new saliency map. These models fall into three categories (Wang et al., 2013) based on how the depth information is used: stereovision models, depth-weighting models, and depth-saliency models. For a comprehensive review of 3D visual attention models, see Wang et al. (2013); Ma and Hang (2015).

While the depth-weighting and depth-saliency models assume that a depth map has been computed, without specifying how, stereovision models explicitly implement

the computation of depth information from the left and right views of the scene, thus replicating the human visual system's stereoscopic perception. An example of this is a study by Bruce and Tsotsos (2005), which extended a 2D selective tuning model of attention to also incorporate binocular information. However, no quantitative assessment of this model was performed.

Depth-weighting models use a base 2D saliency model (computed using one of the existing methods) and then multiplicatively weight the resulting saliency map with the depth information. Regions that are closer to the observer obtain higher weights, corresponding to greater combined saliency. In a model developed by Lang et al. (2012), novel depth priors are learned from a training portion of the data, and these are then combined with the output of a 2D saliency model either using pixel-wise addition or multiplication. With these depth priors, the authors find an increase of performance by 6-7% on their dataset compared to the base 2D model without depth information.

Depth-saliency models come in two flavors. In one, both a depth saliency map, obtained from depth alone, and a more traditional saliency map, obtained from 2D information alone, are computed. The two maps are then linearly combined to generate the final saliency map. Wang et al. (2013) determine depth saliency in a separate experiment involving synthetic stereoscopic stimuli, which allows them to reduce the influence of monocular depth cues, as well as control for the depth of objects and the background. With their experimental results, they propose a probabilistic model of

depth saliency, where the probability of a point being fixated in 3D space is related to the magnitude of center-surround differences in depth contrast. Linearly combining these two saliency maps in a 1:1 ratio (50% weight each for 2D features and depth information) results in better performance on their dataset. In the second type of depth-saliency models, depth information is treated as an additional feature channel, on the same footing as intensity, color, orientation *etc.* The final saliency map is then a function of depth as well as of these other features (Ouerhani and Hügli, 2000; Jost et al., 2004; Hügli et al., 2005).

Our approach falls in the latter class of depth-saliency models, where all image features, including depth, interact through linear combination resulting in the final saliency map. Our model is completely integrated – depth information is treated as another cue which interacts with 2D Gestalt cues to influence figure-ground assignment of proto-objects within the scene. This agrees with anatomical and neurophysiological data that show that disparity selective cells, which are important for encoding stereoscopic depth information, are found in the same early cortical areas as neurons representing other features used in typical saliency models, like color and orientation (Hubel and Wiesel, 1962; Poggio et al., 1988). Different from previous models (Ouerhani and Hügli, 2000; Jost et al., 2004; Hügli et al., 2005), our model is not only based on basic image features (like color, intensity *etc.*) but it includes elements of perceptual organization, in particular proto-objects. The model is an extension of a previously described 2D model (Russell et al., 2014) and is constructed by

including depth information as an additional feature. All features are used to determine proto-object based saliency. We tested our model on the three 3D eye tracking datasets listed in the next section, and we compared results with and without the added depth information.

5.2.2 3D eyetracking datasets

A common method for evaluating the quality of computational models of visual attention is to compare their performance in the prediction of human eye movements. Since its introduction by Parkhurst et al. (2002), this method has been used in a large number of studies, both for static and dynamic scenes (video) and both for human and non-human primates (for a recent review see Borji and Itti, 2013). Nearly all of this work has been limited, however, to 2D scenes.

In order to evaluate 3D attention models, eye tracking data on a variety of visual scenes have to be collected. We use datasets of natural images consisting of color images and associated depth maps along with human fixations for each image. Predictions of saliency maps for eye movements can then be compared to the ground truth fixation data using various metrics. Below, three such publicly available datasets are described. Figure 5.1 shows one example of the data available from each of them.

The NUS-3D dataset (Lang et al., 2012) contains 600 RGB and depth image pairs, each with a resolution of 640×480 pixels. The images show various scenes around the National University of Singapore (NUS) campus and were collected with a

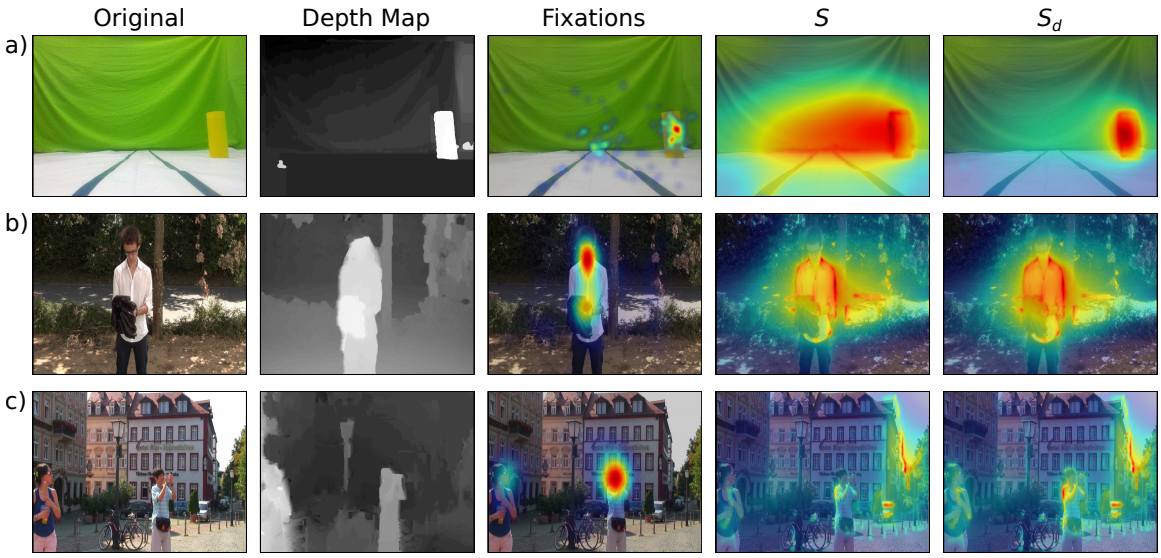


Figure 5.1: Examples of data used and results obtained. Columns (left to right) show one example each of the original image with its corresponding depth map, fixation map, our saliency model without (S) and with (S_d) depth information for the three 3D eye tracking datasets: a) NUS-3D. b) Gaze-3D. c) NCTU-3D.

Microsoft Kinect camera, which is capable of recording both RGB and depth images. The Kinect depth sensor is affected by ambient lighting and has a depth range of only about 4 m, which restricts the types of scenes it can accurately capture. The images were presented to 80 participants and each participant's eye tracking data was captured in both 2D and 3D free-viewing experiments. The 3D stimuli were generated by virtual view synthesis (see Lang et al., 2012, for details) but the synthesized 3D images are not available to the public. Only the raw and smoothed depth maps from the Kinect as well as the fixation density maps for the 2D and 3D viewing conditions are available.

The Gaze-3D dataset (Wang et al., 2013) consists of 18 stereoscopic images, along with their associated disparity maps and perceived depth maps (perceived depth

is computed from raw disparity by taking into account viewing distance and display properties; see Wang et al., 2013, for details). The ground truth disparity maps were calculated from separate left and right image views using an optical flow method (Werlberger et al., 2009). The images come from the Middlebury 2005/2006 stereo image dataset (Scharstein and Pal, 2007) and the IVC 3D image dataset (Urvoy et al., 2012). The dataset also contains raw eye tracking data for both the left and right eyes, as well as processed fixation density maps from a total of 35 participants. The images are high resolution (1300×1080 pixels for the Middlebury subset and 1920×1080 pixels for the IVC subset) and have relatively accurate depth information. A limitation is the small number of images in the dataset.

The NCTU-3D dataset (Ma and Hang, 2015) consists of 475 2D images with a resolution of 1920×1080 pixels along with their corresponding depth maps and eye tracking data. The eye tracking data is in the form of fixation density maps and binary fixation maps. The images in the dataset were collected from randomly selected frames extracted from 11 different sequences of 3D videos from either Youtube (youtube.com) or 3dtv (3dtv.at). The depth maps were generated from left and right eye views using Depth Estimation Reference Software (DERS; version 5.0). A total of 16 participants freely viewed the videos in 3D.

5.3 Methods

5.3.1 Model

Our approach is based on the proto-object based saliency model proposed by Russell et al. (2014). In the model, grouping cells group visual features into proto-objects that are characterized by their locations and spatial scales. The large annular receptive fields of the grouping cells enforce the Gestalt principles of closure and convexity, which in turn biases the activity of proto-objects towards the center of objects. Proto-objects are then a means to organize the scene into separate figures as well as the background. The grouping mechanism operates on multiple feature channels and incorporates competition between proto-objects of similar size and feature type. The model explains the development of border-ownership findings in primate cortex (Craft et al., 2007; Zhou et al., 2000). Given that human eye movements tend to fall predominantly on objects (Einhäuser et al., 2008, but see Borji et al. (2013) for a different view), which are often closed and convex, the locations of proto-objects are also assumed to correlate with the salient points within the image. For a full description of the original model, we refer the reader to Russell et al. (2014).

We extend this model to include depth information. Since some images have border artifacts, all images are cropped to avoid spurious model responses at the borders. To achieve scale invariance, we create an image pyramid spanning 8-10 octaves (depending on the size of the image) by successively down-sampling the input

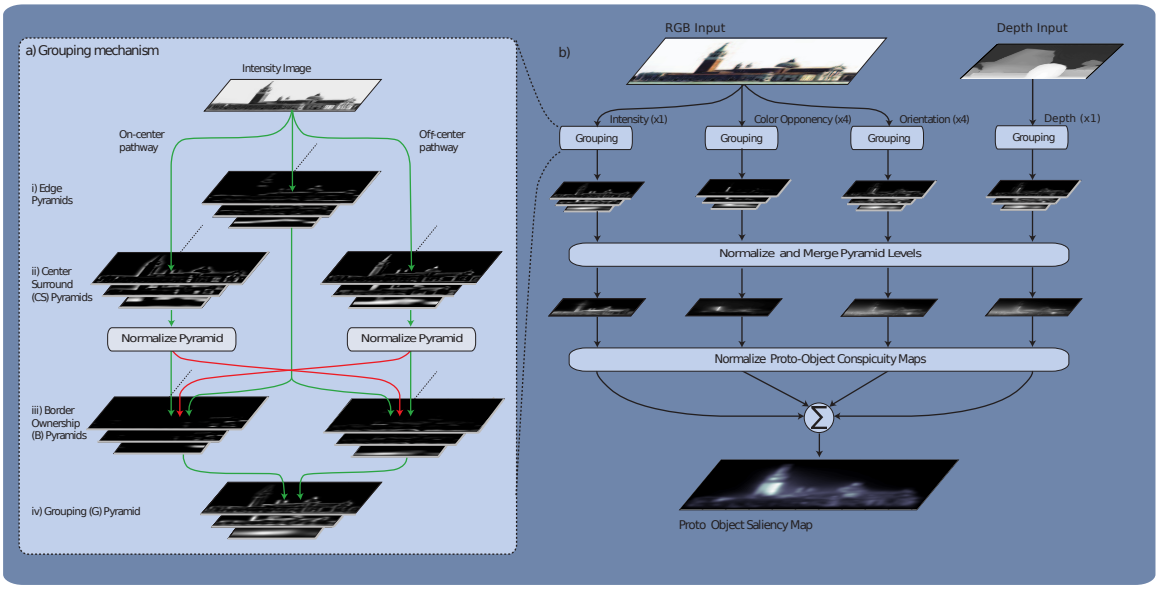


Figure 5.2: Proto-object saliency model with added depth information. The depth map is represented by the image at the top, far right, and the 2D image is to its left. Based on Figure 5 of Russell et al. (2014).

image in steps of 2. We use a minimum cut-off image size of 3×3 pixels, such that pyramid levels that would reduce the image size below 3×3 pixels are not included in our model. All operations are applied independently to each feature and at each level of the feature pyramids, except for when the scales and features are combined to obtain the final saliency map. Each layer of the network represents the neural activity of an array of neurons which tile the visual scene and which are propagated to other layers in the network through feedforward connections. The receptive fields of neurons are described by different correlation kernels, and the image input to each neuron in the model is calculated using the correlation operation. The model was implemented using MATLAB (Mathworks, Natick, MA). A model overview is shown in Figure 5.2.

In this chapter, we will refer to the saliency map generated by the original 2D model without depth information as S, and to the saliency model generated by our new model, which incorporates depth information as S_d. To compute S, the model accepts an input RGB image and decomposes it into different feature channels: one intensity channel, four color-opponency channels, and four orientation channels. Rather than providing "raw" stereoscopic disparity information in the form of different images to the two eyes (or retinae), we assume that the transformation from disparity information to depth has been performed at the input level to our model. There is well-known neuronal circuitry that transforms stereoscopic disparity into depth information (*e.g.* Poggio and Poggio, 1984) and the explicit representation of depth is more appropriate for the intermediate-vision conceptual level of our proto-object based model than the "raw" representation in terms of binocular disparity. Therefore, to compute S_d, we add an additional depth feature channel obtained from the input depth image. Within each feature channel, we perform edge filtering (using oriented Gabor filters at 4 orientations) to obtain the location of proto-object borders. In order to perform feedforward computation of figure-ground segregation, we employ a center-surround mechanism, similar to that used by Itti et al. (1998); such mechanisms have been observed at multiple stages in the brain, including retina, lateral geniculate nucleus, and cortex (*ibid*). This center-surround mechanism provides context information about proto-objects, and biases the activity of border-ownership cells with preferred directions that match the location of the figure.

For the 2D features used in our model, the center-surround mechanism is symmetric with respect to figure-ground contrast polarity (*e.g.* light figures on dark backgrounds or dark figures on light background result in the same net salience contribution). In contrast, for the depth channel we compute the center-surround differences in an asymmetrical manner, consistent with the kind of information provided by stereoscopic depth. While most feature differences across a contour are not predictive about which of its sides is the foreground[1], stereoscopic depth (disparity) provides nearly unambiguous information about which side of the border is closer to the observer. This side "owns" the object border when considered in a depth ordering sense and is part of the foreground. Physiological data show that the responses of border-ownership cells to disparity differences across a figure edge are in agreement with this observation (Zhou et al., 2000; Qiu and von der Heydt, 2005). Therefore, in our model near *vs.* far depth differences bias the activity of border-ownership cells such that the near side is more likely to be classified as the foreground object. Integrating depth information into the representation of an object by its contours is a critical difference between our model and other depth saliency models which directly combine depth feature information with 2D information, without taking into account perceptual grouping effects (*e.g.* Ouerhani and Hügli, 2000; Jost et al., 2004; Hügli et al., 2005). By modeling the response of border-ownership cells to depth edges, we enforce an additional constraint on how depth information is to be combined with 2D

[1]Exceptions are T-junctions (Heitger and von der Heydt, 1993) and extremal edges (Palmer and Ghose, 2008; Ramenahalli et al., 2011, 2012, 2014) both of which are local cues that provide information about edge polarity.

information in order to produce proto-objects and the resulting salient points of the image.

Each channel is processed independently of the others by the same grouping mechanism, and then combined through a series of normalization operators, which allow for competition between proto-objects of similar scale and feature. The final feature map is obtained by scaling each map to a common scale (approximately the middle level of the image pyramid), and then performing a pixel-wise addition across scales. The different feature conspicuity maps are then normalized again and linearly combined with equal weights to form the proto-object saliency map S. When depth information is used, we use a linear combination of 80% 2D features, split equally among intensity, color, and orientation, and 20% depth features to form the depth-added proto-object saliency map S_d. Obviously, this fraction can be modified but we found that results do not critically depend on its choice (see also the Discussion).

5.3.2 Evaluation metrics

We evaluate the performance of our model by comparing our generated saliency maps with ground truth data in the form of human eye fixations. Different from many other studies, we tested our model with a whole battery of metrics that were *not* necessarily chosen to give good results. We did this in order to provide a more diverse picture of the model's overall performance. Riche et al. (2013) have suggested that at least three different metrics are needed to fairly evaluate a given model. To

ensure that our results do not depend on the use of a specific evaluation method, we use a battery of commonly used saliency metrics: area under the curve (AUC), Pearson's linear cross-correlation (PLCC), normalized scanpath saliency (NSS), similarity (SIM), earth-mover's distance (EMD), and the Kullback-Leibler divergence (KLD). For a recent review see Riche et al. (2013), the following is a brief description of the metrics.

KLD, EMD, PLCC, and SIM are distribution-based metrics that measure the similarity/dissimilarity between two distributions (in our case, between the distribution of human eye fixations and of the salient points as predicted by the model). Larger values of KLD and EMD indicate a larger overall difference between the two distributions, while a value of zero indicates that the two distributions are not systematically different from each other. PLCC and SIM are bounded values, where a value of unity indicates that the two distributions are identical, while a value of zero indicates that the distributions are completely uncorrelated (PLCC can also be negative, indicating a negative correlation between the two distributions). AUC is a location-based metric, a measure borrowed from signal-detection theory. An equal number of fixated and random pixels are first chosen from the saliency map. A threshold is then applied to the saliency map, which acts as a classifier, with all saliency points above threshold considered "fixated," and all saliency points below threshold considered "background." For each threshold value, we can then determine a true positive rate and a false positive rate based on the ground truth eye fixation map,

which allows us to generate a Receiver-Operator Characteristic (ROC) curve and calculate the corresponding Area Under the Curve (AUC) metric. An ideal score is unity while a random classification gives a score of 0.5 and systematic mis-classifications result in values between 0 and 0.5. NSS is a value-based metric, which compares predicted saliency values with the corresponding eye fixation maps. NSS effectively measures the average number of standard deviations that the predicted salient points are above the global mean of the saliency map, with larger values indicating fixated points having a higher saliency as predicted by the model.

With the exception of the KLD metric, the code for all evaluation metrics can be found online on the MIT Saliency Benchmark webpage (Judd et al., 2012). The metrics compare the saliency map with either the binary fixation maps that contain the locations of all eye fixations of all participants without smoothing, or the continuous fixation density maps (smoothed averages of fixations). For the datasets we used, both continuous and binary fixation data were either included with the dataset, could be generated from the raw eye tracking data, or were obtained through correspondence with the authors that collected the data. Fixation density maps were used with the PLCC, SIM, KLD, and EMD metrics and binary fixation maps with the AUC and NSS metrics.

To determine whether the addition of depth information improves performance of the base 2D saliency model, we performed two-tailed, paired Student t-tests, with a significance level of $\alpha = 0.05$. To adjust for multiple comparisons and the dependence

between saliency metrics, we applied a Benjamini-Hochberg correction (Benjamini and Hochberg, 1995) to control the false discovery rate ($q = 0.05$). Table 5.1 summarizes the results of our model with and without depth information for the different 3D eye tracking datasets. We also report adjusted p-values in the table.

5.4 Results

For all three datasets, adding depth information ("2D model + depth" compared to "2D model only" in Table 5.1) improved the model's prediction of perceptual saliency in terms of eye fixations. At least three, and sometimes more of the six metrics with and without depth information differed in a statistically significant manner ($p < 0.05$ for the NUS-3D and NCTU-3D datasets, and $p < 0.10$ for the Gaze-3D dataset), although which of the metrics reached significance varied between datasets.

For the NUS-3D dataset, adding depth information improved the PLCC, SIM, AUC, and NSS metrics for both the 2D and 3D viewing conditions ($p < 0.05$, see Table 5.1 for the associated test statistics and p-Values). The EMD and KLD metrics showed improvement that was not statistically significant or no improvement, respectively. For the Gaze-3D dataset, adding depth information improved each of the metrics, but this improvement was not statistically significant at the chosen alpha level ($p > 0.05$). We note here that our model outperforms a previous model (in terms of the PLCC, AUC, and KLD metrics) that was evaluated using the same

Eyetracking Dataset		Saliency Metrics					
		PLCC	SIM	AUC	NSS	EMD	KLD
NUS-3D	2D model only (2D Fixations)	0.349	0.305	0.769	1.036	2.917	1.485
	2D model + depth (2D Fixations)	**0.359****	**0.307****	**0.774****	**1.068****	2.913	1.485
	t(df)	6.58(599)	3.75(599)	4.99(599)	6.75(599)	-0.40(599)	0.02(599)
	p-Value	3.00×10^{-10}	2.87×10^{-4}	7.10×10^{-6}	2.04×10^{-10}	0.831	0.988
	2D model only (3D Fixations)	0.336	0.289	0.772	1.046	2.987	1.559
	2D model + depth (3D Fixations)	**0.347****	**0.291****	**0.777****	**1.080****	2.980	1.559
	t(df)	6.75(599)	4.20(599)	5.27(599)	7.08(599)	-0.68(599)	-0.33(599)
	p-Value	1.04×10^{-10}	4.65×10^{-5}	2.05×10^{-7}	2.50×10^{-11}	0.600	0.742
Gaze-3D	2D model only	0.535	0.682	0.699	0.720	2.108	0.327
	2D model + depth	0.552*	0.688*	0.705	0.743	2.060	0.312*
	t(df)	2.18(17)	2.74(17)	1.79(17)	1.53(17)	-0.98(17)	-2.36(17)
	p-Value	0.086	0.084	0.137	0.174	0.342	0.086
NCTU-3D	2D model only	0.473	0.513	0.760	1.052	3.751	0.755
	2D model + depth	**0.479****	0.514	**0.764****	**1.071****	3.761	0.755
	t(df)	2.35(474)	1.04(474)	2.84(474)	3.1304(474)	0.76(474)	$<10^{-3}$(474)
	p-Value	0.038	0.446	0.014	0.011	0.540	0.999

Table 5.1: Depth information improves saliency prediction on 3D eye tracking datasets. A double asterisk (**) and **boldface type** indicate that the performance of the model with depth information differs significantly from that of the corresponding 2D model, in the row immediately above it (paired t-test with Benjamini-Hochberg correction for multiple comparisons, $p < 0.05$). Similarly, a asterisk (*) indicates that the performance of the model with and without depth information differs significantly, but at a higher alpha level ($p < 0.10$). The value of the t-test statistic (t), the degrees of freedom (df), and the adjusted p-values (p-Value) are also reported here.

dataset (Wang et al., 2013). We also note that at the higher significance level used in that study, the improvement in the PLCC, SIM, and KLD metrics are statistically significant ($p < 0.10$). For the NCTU-3D dataset, adding depth information significantly improved the PLCC, AUC, and NSS metrics ($p < 0.05$), and also improved

the SIM and KLD metrics but, again, these did not reach statistical significance. The EMD metric increased with depth information, indicating a greater difference between the distribution of salient points predicted by the model and the eye fixations, but this difference was not significant.

A special case is the NUS-3D dataset for which eye tracking data from a 2D viewing condition is also available. In this case, participants viewed the images binocularly without stereoscopic depth cues (identical input presented to both eyes) but monocular cues (like occlusion, shading, extremal edges, T-junctions *etc.*) remained available. While depth information plays an important role in the computation of proto-object representations in our model, it does not take into account other monocular depth cues. By comparing the fixation prediction performance of the model when it has access to depth information compared to when it does not (the first and second line of Table 5.1, respectively), we can assess the importance of monocular cues not included in the model. Results reveal significant differences for four of the six metrics (PLCC, SIM, AUC, and NSS, $p < 0.05$). As a result, we conclude that monocular depth cues play an important role for saliency prediction and that future models will likely benefit from including their influence.

Overall, incorporating depth into our proto-object based saliency model improved performance across all three tested datasets, as measured by different metrics that are sensitive to different components of the data. It should be noted that the effect of adding depth information is relatively small, which may point to the relative impor-

tance of traditional 2D features in visual saliency. In our model, depth information generally helps with perceptual saliency prediction, although the degree to which it does may vary greatly based on image content. We provide additional reasons for why the absolute size of the effect is small in the Discussion.

5.5 Discussion

5.5.1 Comparison to previous models

For all 3D datasets, we found that incorporating binocular depth information in our model resulted in a small, but statistically significant improvement in perceptual saliency prediction on most of the evaluation metrics.

For the NUS-3D saliency dataset, adding depth information improved performance of the proto-object based saliency model for both 2D and 3D viewing conditions. The results for the 2D viewing condition agree with the previous finding that incorporating depth information gained from monocular depth cues can improve 2D saliency prediction (Ramenahalli and Niebur, 2013). In that study, however, depth information was inferred from the 2D image using the Make3D algorithm (Saxena et al., 2009), which computes a depth map from a 2D image, while in the current work, depth information is directly collected with the Kinect sensor.

Although our model performance does not exceed that of previously reported results on the NUS-3D dataset (Lang et al., 2012) or the NCTU-3D dataset (Ma and

Hang, 2015), our model has the advantage of being a straightforward extension of an existing 2D model (Russell et al., 2014) which is based on biologically-realistic features of early and intermediate primate vision. Importantly, different from previous work, our model does not rely on learning novel depth priors or, for that matter, learning *anything* from a training set of images. This has at least two advantages. First, we eliminate the time and computational effort needed for training, which typically scales with the number of images and/or the number of features chosen to be learned. Second, using depth information in a way that combines 2D Gestalt cues with depth cues is a mechanism of general validity, and therefore we believe that our model is applicable to a wide range of natural images, not just those included in the training datasets, or images similar to those. We also note that our model does extremely well on the Gaze-3D dataset, significantly outperforming the best previously reported results. This indicates that the proto-object based saliency model may be able to capture perceptual saliency more successfully than other 2D saliency models that are purely feature-based. However, we note that model performance even without the depth information is very good. While depth information does help in the prediction of eye fixations, its contribution is relatively small compared to that from 2D features and not statistically significant at the chosen alpha level ($\alpha = 0.05$).

5.5.2 Relative contribution of 3D features

We combined the 2D and 3D features with a 4:1 ratio of 2D features to 3D features, meaning a weight of 20% on the depth information and of 80% on the traditional 2D features. In contrast, Wang et al. (2013) used a ratio of 1:1 for weighting 2D features and depth information, giving depth information the same importance as the combination of all 2D features. In our experience, at least for conditions under which the three data sets that we have access to were collected, the contribution of depth information is comparable to some of the 2D submodalities, but substantially smaller than the combination of *all* 2D submodalities. Informal parametric studies showed that although the assignment of detailed relative weights is not critical, a clear dominance of 2D over depth information gave the best results.

The performance enhancement due to adding the depth channel results is significant but its absolute value is small. One reason for the small size of the effect could be the long viewing times which were, for the three datasets used, in the range of 4-15 seconds. With these relatively long viewing times, 2D features may play a more critical role in directing the participants's gaze, compared to the role of 3D features, which may be more important early in the viewing period. Indeed, others have shown time-dependent influences of the 2D and 3D features on saliency prediction (Gautier and Le Meur, 2012).

Secondly, we did not divide our images based on the depth range or depth-of-field, which have been shown to be important factors in determining to what extent

depth information can influence visual attention (Lang et al., 2012). It is possible that large depth differences are disproportionately salient, but large differences can only occur in images with a large depth of field. Similarly, depth information may be particularly advantageous in highly-textured scenes, where 2D cues are not enough to perform segmentation of proto-objects. Ma and Hang (2015) present examples of images where depth information may help participants to segment objects among highly-textured backgrounds, or among objects that are not located at the center of the image.

A third reason why the absolute size of the 3D effect is small may be the presence of common surfaces in the images, which can influence the perception of metric depth. Several previous psychophysical studies have shown that perceived binocular depth can be affected by a common surface (McKee, 1983; Glennerster and McKee, 1999; He and Ooi, 2000). Importantly, a recent result shows that perceived absolute distance to objects on the ground surface is not different between monocular and binocular viewing conditions (Ooi and He, 2015). Binocular disparity may only play a critical role in perceiving absolute distance of a target in midair. Given that many objects rest on surfaces, binocular disparity information may not be needed – the relative depth of objects along a perceived common surface under 2D viewing conditions may be sufficient. It is then possible that the small absolute size of the 3D effect is due to the fact that many of the images in the datasets are natural scenes that carry common surfaces, such as the ground, floor, walls *etc.* (see Figure 5.1). This provides further

evidence for the important role of surfaces defined either monocularly or binocularly for the perceptual organization of visual scenes (He and Nakayama, 1992, 1995; Hu et al., 2015).

Our proto-object based model operates at intermediate stages of vision, where perceptual organization of the visual scene is thought to occur. In contrast, other 3D saliency models use only low-level visual features, or incorporate additional semantically important cues that may only be found in higher visual areas. These other models either treat depth as an early feature which can be combined multiplicatively or additively with the 2D saliency map, or they include other features (such as human face and body detection, Cerf et al., 2008) as a means to improve performance. We believe that while adding these other features can improve model performance in many cases, intermediate stages of vision are critical for transforming low-level visual features into higher-level object representations that form the basis for further visual processing and allocation of attention. We show that by incorporating depth information as an additional cue into the grouping mechanism, we can more accurately predict where participants will fixate within a scene (*i.e.* as a marker of perceptual saliency). This is because binocular depth provides unambiguous information about the location and border ownership of object edges, which can be used for the perceptual organization of the scene in terms of proto-objects.

5.5.3 Object-based saliency

There has been some debate as to whether the computation of visual saliency is feature-based or object-based. Feature-based models rely on low-level feature contrast to generate a saliency map (*e.g.* Itti et al., 1998; Walther and Koch, 2006). Object-based saliency models, instead, start from the assumption that objects, and not necessarily their constituent features, are what is needed for determining the salient regions of an image and are the primary driver of fixations (Einhäuser et al., 2008; Nuthmann and Henderson, 2010; Stoll et al., 2015). Support for object-based models comes from the analysis of fixation locations within objects. Fixations are well described by a two-dimensional Gaussian distribution with a mean biased towards the center of the object, which represents the preferred viewing location (PVL) of the object (Nuthmann and Henderson, 2010). Critically, proto-objects computed solely in terms of low-level features without the influence of Gestalt cues (Walther and Koch, 2006) do not exhibit a central PVL. However, the proto-objects in by our model integrate low-level feature information from different spatial locations and scales, such that their final activity should be biased towards the center of closed, convex objects. As a result, previous experimental results that cast doubt on the role of feature-based saliency (Einhäuser et al., 2008; Nuthmann and Henderson, 2010; Stoll et al., 2015) do not rule out our proto-object based model, but rather support it. A direct comparison of the proto-objects in our model with real objects is still lacking, so it remains to be seen whether these proto-objects exhibit a central PVL,

which is an area of future research. We believe our model fits most closely with the definition of proto-objects by Rensink (2000), as providing both a feedforward measure of objecthood and a "handle" for top-down processes. Saliency is then a function of proto-objects, and proto-objects may also causally drive attention. For a more detailed discussion of this issue, we refer the reader to Russell et al. (2014).

5.6 Conclusion

We introduce a new proto-object based saliency model which makes use of information about 3D depth to segment natural scenes. Our model is an extension of previous models in 2D and it is constructed from first principles, without relying on learning of depth priors or depth features; it does not require any training. The model is biologically-inspired, with the computations needed being directly mapped to neural mechanisms that have been found in the brain. Using data from three separate 3D eye tracking datasets, we show that depth information improves performance in a robust manner using a number of evaluation metrics. Athough we find that proto-objects are largely formed based on 2D features, the added depth information has clear benefits in improving performance of the model in terms of predicting the location of human eye fixations.

Chapter 6

Conclusion

Visual processing of objects makes use of both feedforward and feedback streams of information. In this thesis, we showed how recurrent neural networks that make use of grouping mechanisms can perform a wide variety of tasks, including contour integration, figure-ground segregation (both in artificial and natural scenes), grouping of 3D surfaces, and visual saliency prediction. A key feature of our models is the use of grouping neurons whose activity represents tentative objects (proto-objects) based on the integration of local feature information. Grouping neurons receive input from an organized set of local feature neurons, and project modulatory feedback to those same neurons. Additionally, inhibition at both the local feature level and the object representation level biases the interpretation of the visual scene in agreement with principles from Gestalt psychology. Our models can explains several sets of psychophysical and neurophysiological results (He and Nakayama, 1995; Zhou et al.,

2000; Qiu et al., 2007; Chen et al., 2014; Williford and von der Heydt, 2016), and makes testable predictions about the influence of neuronal feedback and attentional selection on neural responses across different visual areas. Our proposed models also provide a framework for understanding how object-based attention is able to select both objects and the features associated with them.

Future work will need to further explore how grouping mechanisms operate in 3D. Understanding how the brain efficiently represents 3D surfaces, possibly with the combination of basis functions and grouping mechanisms, will be critical for advancing our understanding of the cortical mechanisms of perceptual organization. Currently, our models can only handle static images. Additional work on how grouping is performed on other feature types, such as motion, and how information from different features is ultimately integrated into a whole, will be important for extending our models to handle the rich spatiotemporal information present in videos. Finally, the interaction between mid-level visual representations of proto-objects and higher-level visual representations of object identity is still unknown. Understanding how object identity influences figure-ground segmentation and other grouping processes is a promising area of future research.

Appendix A

Grouping Model Network

A.1 Details of model implementation

One pixel in our model input represents the typical size of the receptive field for a V1 edge cell (which at eccentricity 5 deg is about 0.7 deg, Chen *et al.* 2014). The simulated visual field is assumed to be homogeneous (*i.e.* , we neglect the influence of the cortical magnification factor). To avoid unbalanced inputs near the boundaries of the visual field we use periodic boundary conditions. Each neuronal unit in the model represents multiple neurons with overlapping receptive fields and similar tuning. It is referred to as a "neuron" in the following and it is represented by an ordinary differential equation, eq. 1. For all examples the system is simulated for 0.5s and an equilibrium is reached within a few tens or about hundred ms. A typical simulation of a visual field of 64×64 units consists of 30,528 coupled ordinary differential equations

which are solved using a fourth-order Runge-Kutta algorithm in MATLAB.

Each neuron is characterized by its spatial location, its type (edge cell, object grouping cell, *etc.*) and one additional property, as follows. E cells are indexed by the angle of their preferred orientation: 0, $\pi/4$, $\pi/2$, and $3\pi/4$ (all angles relative to the horizontal). B cells are indexed by the angle of their preferred side of figure: right, 0; upper right, $\pi/4$; up, $\pi/2$; upper left, $3\pi/4$; left, π; lower left, $5\pi/4$; down, $3\pi/2$; lower right, $7\pi/4$. Note that the preferred side of a B cell is always orthogonal to its preferred orientation. Contour grouping cells are indexed by their preferred orientation (0, $\pi/4$, $\pi/2$, and $3\pi/4$), while object grouping neurons are selective for annuli of a preferred radius. As discussed earlier, we only use one preferred radius in this model, which we chose as corresponding to 16 pixels in the input layer. The network receives two types of inputs. A binary edge map activates E cells, taking value 1 if an edge of that orientation is present at this location in the image and (-1) otherwise. Finally, an attentional field stimulates grouping cells, with maximal strength of 0.07. This value is listed in Table A.1, as are those of all other model parameters.

In the following, we define the anatomical connection patterns between all neuronal populations in our model. Obviously, this defines the receptive (and projective) fields of all model neurons. At the input level, edge cells (E population) receive one-to-one connections from input units (IN population) of the respective preferred orientations o_i at position (x, y), resulting in a connectivity weight distribution as

follows,

$$WIN^{o_1}_{x,y}E^{o_2}_{x,y} = \begin{cases} intoe_w & if\ o_1 = o_2 \\ -1 & otherwise \end{cases} \tag{A.1}$$

where the weight of the input to edge cell connections, $intoe_w$ is a scaling parameter, and its value is set to 1. Changing it does not produce different activation patterns in the network but rather scales up all activities.

The connections from edge cells to inhibitory IE cells are two-dimensional isotropic Gaussians:

$$WE_{x_1,y_1}IE_{x_2,y_2} = N_{etoie}\ \exp\left(-\frac{(x_1-x_2)^2}{2\,etoie^2_{sd}} - \frac{(y_1-y_2)^2}{2\,etoie^2_{sd}}\right) \tag{A.2}$$

where the normalization coefficient N_{etoie} is determined from:

$$etoie_w = \sum_{i,j=-24}^{24} WE_{x,y}IE_{x+i,y+j} \tag{A.3}$$

The weight of the edge to IE connections $etoie_w$ is set to 8. The standard deviation of the lateral connections, $etoie_{sd}$ was chosen to be eight times the size of the edge cells' receptive field size. For simplicity, all lateral connections in V1 are assumed to

have the same standard deviation. The upper limit in the sum of eq. A.3 is where we truncate the Gaussian distribution used to define the connectivities, at three times its standard deviation (8), equal to 24 times the size of a V1 RF.

IE cells are not orientation selective, and inhibit all edge cells in their neighborhood. The strength of inhibition is independent of the preferred orientation of the edge cells and has the same pattern as the reciprocal edge to IE cell connections:

$$WIE_{x_1,y_1}E_{x_2,y_2} = N_{ietoe}\ \exp\left(-\frac{(x_1-x_2)^2}{2\,ietoe_{sd}^2} - \frac{(y_1-y_2)^2}{2\,ietoe_{sd}^2}\right) \tag{A.4}$$

where the normalization coefficient N_{ietoe} is determined from:

$$ietoe_w = \sum_{i,j=-24}^{24} WIE_{x,y}E_{x+i,y+j} \tag{A.5}$$

The strength of the inhibitory connections IE to edge cells, $ietoe_w$, is an important parameter in the model. Its value was chosen to be -8, which is just strong enough for the inhibition of the edge cells to cancel out activity in the case of a uniform stimulation field. In our model, attention in the absence of edges is a broad field, and experimental observations show little effect of attention alone on the firing rates of purely sensory (edge) cells. E cells locally connect to other edge cells of the same preferred orientation. These connections allow for the passing of contour information

along a line. The connection weights are,

$$
\begin{aligned}
WE^{o_1}_{x_1,y_1}E^{o_2}_{x_2,y_2} &= N_{etoe}\,\delta_{o_1,o_2}\,\delta_{y_1,0}\,\delta_{y_2,0}\,\exp\left(-\frac{(x_1-x_2)^2}{2\,etoe^2_{sd}}\right) \text{ if } o_1,o_2=0\\
WE^{o_1}_{x_1,y_1}E^{o_2}_{x_2,y_2} &= N_{etoe}\,\delta_{o_1,o_2}\,\delta_{x_1,y_1}\,\delta_{x_2,y_2}\,\exp\left(-\frac{(x_1-x_2)^2}{2\,etoe^2_{sd}}\right) \text{ if } o_1,o_2=\pi/4\\
WE^{o_1}_{x_1,y_1}E^{o_2}_{x_2,y_2} &= N_{etoe}\,\delta_{o_1,o_2}\,\delta_{x_1,0}\,\delta_{x_2,0}\,\exp\left(-\frac{(y_1-y_2)^2}{2\,etoe^2_{sd}}\right) \text{ if } o_1,o_2=\pi/2\\
WE^{o_1}_{x_1,y_1}E^{o_2}_{x_2,y_2} &= N_{etoe}\,\delta_{o_1,o_2}\,\delta_{x_1,-y_1}\,\delta_{x_2,-y_2}\,\exp\left(-\frac{(y_1-y_2)^2}{2\,etoe^2_{sd}}\right) \text{ if } o_1,o_2=3\pi/4
\end{aligned}
\tag{A.6}
$$

where the normalization coefficient N_{etoe} is determined from:

$$
etoe_w = \sum_{i,j=-24}^{24} WE^{o}_{x,y}E^{o}_{x+i,y+j} \tag{A.7}
$$

The weight of the excitatory lateral connections $etoe_w = 2/3$ is chosen such that individual excitatory connections are stronger than the nonspecific inhibitory connections for a line of cells along the preferred direction, and that the integral of the excitatory connections is smaller than the integral of the inhibitory ones. The standard deviation $etoe_{sd} = 8$ is the same as for all other lateral connections in V1.

The edge cells E excite a pair of border-ownership cells with the same preferred orientation o_1 and at the same position, (x_i, y_i). Border ownership selective cells have opposite side-of-figure preferences o_2 which are orthogonal to o_1 (*i.e.* differing by $\pi/2$ in either direction). They are generated by connections from E cells whose weight is

given by a 2D Gaussian,

$$WE^{o_1}_{x_1,y_1}B^{o_2}_{x_2,y_2} = etob_w \exp\left(-\frac{(x_1-x_2)^2}{2\,etob^2_{sd}} - \frac{(y_1-y_2)^2}{2\,etob^2_{sd}}\right)$$
$$o_1 \in \{0, \pi/4, \pi/2, 3\pi/4\}$$
$$o_2 \in \{0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4\ \} \tag{A.8}$$

where the weight of the E to border-ownership connections, $etob_w$ is set to 1.

The connections from border-ownership to inhibitory IB cells are two-dimensional Gaussians:

$$WB_{x_1,y_1}IB_{x_2,y_2} = N_{btoib} \exp\left(-\frac{(x_1-x_2)^2}{2\,btoib^2_{sd}} - \frac{(y_1-y_2)^2}{2\,btoib^2_{sd}}\right) \tag{A.9}$$

where the normalization coefficient N_{btoib} is determined from:

$$btoib_w = \sum_{i,j=-12}^{12} WB_{x,y}IB_{x+i,y+j} \tag{A.10}$$

The weight of the border-ownership to IB connections $btoib_w$ is set to 2. The standard deviation of the lateral connections, $btoib_{sd}$ was chosen to be four times the size of the border-ownership cells' receptive field size. For simplicity, all lateral connections in V2 are assumed to have the same standard deviation.

IB cells, which are not orientation selective, inhibit all border-ownership in their

neighborhood with the same pattern as B to IB cell excitation:

$$WIB_{x_1,y_1}B_{x_2,y_2} = N_{ibtob}\ \exp\left(-\frac{(x_1-x_2)^2}{2\ ibtob_{sd}^2} - \frac{(y_1-y_2)^2}{2\ ibtob_{sd}^2}\right) \tag{A.11}$$

where the normalization coefficients N_{ibtob} is determined from:

$$ibtob_w = \sum_{i,j=-12}^{12} WIB_{x,y}B_{x+i,y+j} \tag{A.12}$$

The strength of the inhibitory connections IB to border-ownership, $ibtob_w$, is an important parameter in the model. Its value was chosen to be -2, which is just strong enough for the inhibition of the border-ownership cells to cancel out activity in the case of a uniform stimulation field. In our model, attention in the absence of edges is a broad field, and experimental observations show little effect of attention alone on the firing rates of border-ownership cells. This parameter also influences the value of the attention modulation of the non-preferred border ownership cells, and the value which was a priori chosen reproduces well the observed experimental value.

B cells connect to other border-ownership cells of the same preferred side of figure.

These connections allow the passing of border ownership signal along a line:

$$
\begin{aligned}
WB^{o1}_{x_1,y_1}B^{o2}_{x_2,y_2} &= N_{btob}\,\delta_{o_1,o_2}\,\delta_{x_1,0}\,\delta_{x_1,0}\,\exp\left(-\frac{(y_1-y_2)^2}{2\,btob^2_{sd}}\right)\;\; o_1,o_2\in\{0,\pi\}\\
WB^{o1}_{x_1,y_1}B^{o2}_{y_2,y_2} &= N_{btob}\,\delta_{o_1,o_2}\,\delta_{x_1,-y_1}\,\delta_{x_2,-y_2}\,\exp\left(-\frac{(y_1-y_2)^2}{2\,etoe^2_{sd}}\right)\;\; o_1,o_2\in\{\pi/4,5\pi/4\}\\
WB^{o1}_{x_1,y_1}B^{o2}_{x_2,y_2} &= N_{btob}\,\delta_{o_1,o_2}\,\delta_{y_1,0}\,\delta_{y_2,0}\,\exp\left(-\frac{(x_1-x_2)^2}{2\,btob^2_{sd}}\right)\;\; o_1,o_2\in\{\pi/2,3\pi/2\}\\
WB^{o1}_{x_1,y_1}B^{o2}_{x_2,y_2} &= N_{btob}\,\delta_{o_1,o_2}\,\delta_{x_1,y_1}\,\delta_{x_2,y_2}\,\exp\left(-\frac{(x_1-x_2)^2}{2\,btob^2_{sd}}\right)\;\; o_1,o_2\in\{3\pi/4,7\pi/4\}
\end{aligned}
\tag{A.13}
$$

where the normalization coefficient N_{btob} is determined from:

$$
btob_w = \sum_{i,j=-12}^{12} WB^{o}_{x,y}B^{o}_{x+i,y+j} \tag{A.14}
$$

The weight of the excitatory lateral connections $btob_w = 2/3$ is chosen such that individual excitatory connections are stronger than the nonspecific inhibitory connections for a line of cells along the preferred direction, and that the integral of the excitatory connections is smaller than the integral of the inhibitory ones. The standard deviation $btob_{sd} = 4$ is the same as for all other lateral connections in V2.

B cells also connect to other border-ownership cells of orthogonal preferred orientation. These connections allow the passing of border ownership information along a corner, where rot is a rotation operator that rotates the connection matrix C_f by the

angle o:

$$\begin{aligned} WB^{o_1}_{x_1,y_1}B^{o_2}_{x_2,y_2} = btob_{wc}\,(&\mathrm{rot}_{o_1}(C_f(x_2-x_1,y_2-y_1))\delta_{o_1,o_2-\pi/2} \\ &+ c\,\mathrm{rot}_{o_1+\pi}(C_f(x_2-x_1,y_2-y_1))\delta_{o_1,o_2-\pi/2} \\ &+ \mathrm{rot}_{o_2}(C_b(x_2-x_1,y_2-y_1))\delta_{o_1,o_2+\pi/2} \\ &+ c\,\mathrm{rot}_{o_2+\pi}(C_b(x_2-x_1,y_2-y_1))\delta_{o_1,o_2+\pi/2}) \end{aligned} \tag{A.15}$$

with the matrices

$$C_f(x,y) = \begin{cases} N_{btobc}\ \exp(-\frac{x^2+y^2}{2\,btob^2_{sd}}) & if\ x \geq 0\ \&\ y < 0 \\ 0 & otherwise \end{cases} \tag{A.16}$$

$$C_b(x,y) = \begin{cases} N_{btobc}\ \exp(-\frac{x^2+y^2}{2\,btob^2_{sd}}) & if\ x \geq 0\ \&\ y > 0 \\ 0 & otherwise \end{cases} \tag{A.17}$$

and the normalization coefficient N_{btobc} is obtained from

$$1 = \sum_{i,j=-12}^{12} C_f(i,j) \tag{A.18}$$

The values of $btob_{wc} = 2/3$ and $c = 2/3$ were chosen to be the same as $btob_w$.

The connections from border-ownership cells to object grouping cells consist of 2D Gaussians rotated to the appropriate angle and arranged in a circular fashion, where

$btogo_{sds}$ and $btog_{sdl}$ are the spreads of the Gaussian in the radial and tangential directions, respectively:

$$WB^{o}_{x_1,y_1}G^{r}_{x_2,y_2} = \mathrm{rot}_o\left(N_{btogr}\ \exp\left(-\frac{(x_1-x_2+r)^2}{2\,(btogo_{sds}r)^2}-\frac{(y_1-y_2)^2}{2\,(btog_{sdl}r)^2}\right)\right)$$
$$o \in \{0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4\} \tag{A.19}$$

where the normalization coefficient N_{btogr} is

$$btog_w = \sum_{i=-2r}^{2r} WB^{0}_{x-r,y+i}G^{r}_{x,y} \tag{A.20}$$

The strength of the border-ownership to grouping connections $btog_w$, is a scaling parameter and was chosen to be 0.125, since in the model there are 4 preferred orientations and two side-of-figure preferences (8 total orientations) for the border-ownership cells, each of which sends input to each grouping cell. Each orientation of border-ownership cells provides a 2D Gaussian input to the grouping cells, with the standard deviation on the direction orthogonal to the radius being 0.5 times the radius, while the standard deviation on the direction parallel to the radius is 0.25 times the radius. (If these standard deviations are too large, then grouping cells lose specificity and, for more complicated images, some edges are assigned incorrectly. If these standard deviations are too small, the density of grouping cells needs to be

increased.) The distance between two neighboring cells of radius r is $r/2$, such that a line is never more than one standard deviation away from the center of a patch of connections involved in its grouping.

The feedback from the object grouping cells to lower level feature-selective neurons follows a similar spatial pattern of the B to object grouping connections. The feedback to E and B are similar, except E cells receive feedback with twice the radius and standard deviations to account for upsampling to twice the number of neurons in the V1 layer:

$$
\begin{aligned}
WG^{r}_{x_2,y_2}B^{o_1}_{x_1,y_1} &= \mathrm{rot}_{o_1}\left(N_{gtobr}\ \exp\left(-\frac{(x_1-x_2+r)^2}{2\,(btogo_{sds}r)^2}-\frac{(y_1-y_2)^2}{2\,(btog_{sdl}r)^2}\right)\right)\\
WG^{r}_{x_2,y_2}E^{o_2}_{x_1,y_1} &= \mathrm{rot}_{o_2+\pi/2}\left(N_{gtoer}\ \exp\left(-\frac{(x_1-x_2+(2r))^2}{2\,(btogo_{sds}(2r))^2}-\frac{(y_1-y_2)^2}{2\,(btog_{sdl}(2r))^2}\right)\right)\\
&\quad+\mathrm{rot}_{o_2+3\pi/2}\left(N_{gtoer}\ \exp\left(-\frac{(x_1-x_2+(2r))^2}{2\,(btogo_{sds}(2r))^2}-\frac{(y_1-y_2)^2}{2\,(btog_{sdl}(2r))^2}\right)\right)\\
o_1 &\in \{0,\pi/4,\pi/2,3\pi/4,\pi,5\pi/4,3\pi/2,7\pi/4\ \}\\
o_2 &\in \{0,\pi/4,\pi/2,3\pi/4\}
\end{aligned}
\tag{A.21}
$$

Since the number of the grouping cells on a line is inverse proportional to their scale, the line integral of the weight is considered proportional to the radius of the G cell:

$$gtob_w = \sum_{i=-2r}^{2r} WG^r_{x,y} B^{\pi}_{x+r,y+i} \tag{A.22}$$

The value of $gtob_w = 2/3$ is critical for the model, and the border ownership modulation index depends critically on it. In order to reproduce the observed attention modulation of the nonpreferred side of figure for B cells, the weight of the feedback to edge cells needs to be four times that to the border ownership cells $gtoe_w = 8/3$.

In previous work (Mihalas et al., 2011), in order to fit the observed reaction time cost observed when irelevant objects are outside the focus of attention, a feedback from G to IG cells is introduced. Here, this connection allows for competition between different grouping neurons. This has the same pattern, but half the scaled weight as the feedback to B cells and twice its standard deviations:

$$WG^r_{x_2,y_2} IG^o_{x_1,y_1} = \mathrm{rot}_o \left(N_{gtobr} \ \exp\left(-\frac{(x_1 - x_2 + r)^2}{2\,(btogo_{sds}(2r))^2} - \frac{(y_1 - y_2)^2}{2\,(btog_{sdl}(2r))^2} \right) \right)$$
$$o \in \{0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4 \ \} \tag{A.23}$$

The connections from inhibitory IG cells to object grouping cells follow a similar connection pattern as the B to G cell excitation, except with the antipreferred

orientation inhibiting the G cells:

$$WIG^{o+\pi}_{x_1,y_1} G^{r}_{x_2,y_2} = \mathrm{rot}_{o+\pi}\left(N_{igtogr}\ \exp\left(-\frac{(x_1 - x_2 + r)^2}{2\,(btogo_{sds} r)^2} - \frac{(y_1 - y_2)^2}{2\,(btog_{sdl} r)^2} \right) \right)$$
$$o \in \{0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4\ \} \tag{A.24}$$

with the normalization coefficient N_{igtogr} obtained from:

$$igtog_w = \sum_{i=-2r}^{2r} WIG^{\pi}_{x-r,y+i} G^{r}_{x,y} \tag{A.25}$$

as well as inhibition from orthogonal orientations:

$$W_o IG^{o\pm\pi/2}_{x_1,y_1} G^{r}_{x_2,y_2} = \mathrm{rot}_{o\pm\pi/2}\left(N_{igtogro}\ \exp\left(-\frac{(x_1 - x_2 + r)^2}{2\,(btogo_{sds} r)^2} - \frac{(y_1 - y_2)^2}{2\,(btog_{sdl} r)^2} \right) \right)$$
$$o \in \{0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4\ \} \tag{A.26}$$

with the normalization coefficient $N_{igtogro}$ obtained from:

$$igtog_{wo} = \sum_{i=-2r}^{2r} W_o IG^{\pi/2}_{x-r,y+i} G^{r}_{x,y} \tag{A.27}$$

It is assumed that the strength of the inhibition from the nonpreferred side of figure $igtog_w$ is equal to that of orthogonal preferred sides of figures $igtog_{wo}$, and

each of them is equal to the value of the excitatory strength $btog_w$. Similar to the lateral connections in V2, this feedback loop has stronger but fewer and more specific excitatory connections, resulting in robust activity for specific inputs and more total inhibitory connection weights, resulting in little activity cased by nonspecific inputs.

Similarly, the connections from border-ownership cells with opposite side of figure preferences to contour grouping cells with the same preferred orientation consist of rotated 2D Gaussians, where $btogc_{sds}$ and $btog_{sdl}$ are the spreads of the Gaussian in the radial and tangential directions, respectively:

$$\begin{aligned} WB^{o_1}_{x_1,y_1} G^{o_2}_{x_2,y_2} &= \mathrm{rot}_{o_1}\left(N_{btogr}\ \exp\left(-\frac{(x_1-x_2)^2}{2\,(btogc_{sds}r)^2}-\frac{(y_1-y_2)^2}{2\,(btog_{sdl}r)^2}\right)\right) \\ &\sin^2(o_1-o_2)=1 \\ o_1 &\in \{0,\pi/4,\pi/2,3\pi/4,\pi,5\pi/4,3\pi/2,7\pi/4\ \} \\ o_2 &\in \{0,\pi/4,\pi/2,3\pi/4\} \end{aligned} \tag{A.28}$$

where the normalization coefficient N_{btogr} is chosen such that

$$btog_w = \sum_{i=-2r}^{2r} WB^0_{x,y+i} G^0_{x,y} \tag{A.29}$$

There is also excitation from other orientations, in order to explain the orientation

dependence of V4 neurons:

$$
\begin{aligned}
W_o B^{o_1}_{x_1,y_1} G^{o_2}_{x_2,y_2} &= \frac{3}{4}\,\mathrm{rot}_{o_1}\left(N_{btogo}\,\exp\left(-\frac{(x_1-x_2)^2}{2\,(btogc_{sds}r)^2}-\frac{(y_1-y_2)^2}{2\,(btog_{sdl}r)^2}\right)\right)\\
&\sin^2(o_1-o_2)=\frac{\sqrt{2}}{2}\\
W_o B^{o_1}_{x_1,y_1} G^{o_2}_{x_2,y_2} &= \frac{1}{4}\,\mathrm{rot}_{o_1}\left(N_{btogo}\,\exp\left(-\frac{(x_1-x_2)^2}{2\,(btogc_{sds}r)^2}-\frac{(y_1-y_2)^2}{2\,(btog_{sdl}r)^2}\right)\right)\\
&\sin^2(o_1-o_2)=0\\
&o_1\in\{0,\pi/4,\pi/2,3\pi/4,\pi,5\pi/4,3\pi/2,7\pi/4\ \}\\
&o_2\in\{0,\pi/4,\pi/2,3\pi/4\}
\end{aligned}
\tag{A.30}
$$

with the normalization coefficient N_{btogo} obtained from:

$$
btog_{wo}=\sum_{i=-2r}^{2r} WB^0_{x,y+i}G^0_{x,y} \tag{A.31}
$$

The strength of the border-ownership to grouping connections $btog_w$, is a scaling parameter and was chosen to be 0.125, since in the model there are 4 preferred orientations and two side-of-figure preferences (8 total orientations) for the border-ownership cells, each of which sends input to each grouping cell. Each orientation of border-ownership cells provides a 2D Gaussian input to the grouping cells, with the standard deviation on the direction orthogonal to the radius being 0.5 times the radius, while the standard deviation on the direction parallel to the radius is 0.1

times the radius. This standard deviation parallel to the radius is smaller than that for object grouping neurons (0.25 times the radius) to account for higher selectivity to contours.

The feedback from the grouping cells also follows the spatial pattern of the B to grouping connections. The feedback to E and B are similar, although E cells receive feedback with twice the standard deviations to account for the higher number of neurons in the V1 layer:

$$
\begin{aligned}
WG^{o_1}_{x_2,y_2} B^{o_2}_{x_1,y_1} &= \mathrm{rot}_{o_2}\left(N_{gtobr}\ \exp\left(-\frac{(x_1-x_2)^2}{2\,(btogc_{sds}r)^2} - \frac{(y_1-y_2)^2}{2\,(btog_{sdl}r)^2} \right) \right) \\
&\sin^2(o_1-o_2) = 1 \\
WG^{o_1}_{x_2,y_2} E^{o_1}_{x_1,y_1} &= \mathrm{rot}_{o_1}\left(N_{gtobr}\ \exp\left(-\frac{(x_1-x_2)^2}{2\,(btogc_{sds}(2r))^2} - \frac{(y_1-y_2)^2}{2\,(btog_{sdl}(2r))^2} \right) \right) \\
&o_1 \in \{0, \pi/4, \pi/2, 3\pi/4\} \\
&o_2 \in \{0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4\ \}
\end{aligned}
\tag{A.32}
$$

Since the number of the grouping cells on a line is inverse proportional to their scale, the line integral of the weight is considered proportional to the radius of the G cell:

$$
gtob_w = \sum_{i=-2r}^{2r} WG^{o}_{x,y} B^{\pi}_{x,y+i} \tag{A.33}
$$

The connection strength to the IE cells residing in V2 is assumed to be the same

as the connection strength to the B cells. The value of $gtob_w = 2/3$ is used for the model, similar to that for the object grouping cell to B cell feedback connections. As for the border ownership results, the weight of the feedback to edge cells is four times that to the border ownership cells $gtob_w = 8/3$.

To fit the observed reaction time cost observed when irelevant objects are outside the focus of attention, a feedback from contour G to IG cells is introduced. This has the same pattern, but half the scaled weight as the feedback to B cells and twice its standard deviations:

$$
\begin{aligned}
WG^{o_1}_{x_2,y_2} IG^{o_2}_{x_1,y_1} = \mathrm{rot}_{o_2} \left(N_{gtobr} \, \exp\left(-\frac{(x_1 - x_2)^2}{2\,(btogo_{sds}(2r))^2} - \frac{(y_1 - y_2)^2}{2\,(btog_{sdl}(2r))^2} \right) \right) \\
\sin^2(o_1 - o_2) = 1 \\
o_1 \in \{0, \pi/4, \pi/2, 3\pi/4\} \\
o_2 \in \{0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4 \,\}
\end{aligned}
\tag{A.34}
$$

The connections from IG to contour grouping are nonspecific and similar to the connections from IG to object grouping cells, except without inhibition from other

orientations:

$$WIG^{o_1}_{x_1,y_1}G^{o_2}_{x_2,y_2} = \mathrm{rot}_{o_1}\left(N_{igtogr}\ \exp\left(-\frac{(x_1-x_2+r)^2}{2\,(btogo_{sds}r)^2}-\frac{(y_1-y_2)^2}{2\,(btog_{sdl}r)^2}\right)\right)$$
$$o_1 \in \{0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4\ \}$$
$$o_2 \in \{0, \pi/4, \pi/2, 3\pi/4\} \tag{A.35}$$

with the normalization coefficient N_{igtogr} obtained from:

$$igtog_w = \sum_{i,j=-3}^{3} WIG^{o_1}_{x,y}G^{o_2}_{x+i,y+j} \tag{A.36}$$

It is assumed that the strength of the inhibition $igtog_w$ is equal to the excitatory strength $btog_w$. Similar to the lateral connections in V2, this feedback loop has stronger but fewer and more specific excitatory connections, resulting in robust activity for specific inputs and more total inhibitory connection weights, resulting in little activity cased by nonspecific inputs.

Table A.1: Parameter Values The scaling parameters with values that do not influence the behavior of the network are marked by *, see text.

Connection	Parameter	Value	Description
input	lE	1	input to edge cells
	$intoe_w$	1^*	input weight
attention	$Gatt_w$	0.07	maximal input
input	$Gatt_{sd}$	8	standard deviation
EtoIE	$etoie_w$	8	total strength
	$etoie_{sd}$	8	standard deviation
IEtoE	$ietoe_w$	-8	total strength
	$ietoe_{sd}$	8	standard deviation
EtoE	$etoe_w$	2/3	total strength
	$etoe_{sd}$	8	standard deviation
EtoB	$etob_w$	1^*	total strength
BtoIB	$btoib_w$	2	total strength
	$btoib_{sd}$	4	standard deviation
IBtoB	$ibtob_w$	-2	total strength
	$ibtob_{sd}$	4	standard deviation
BtoB	$btob_w$	2/3	total strength
	$btob_{wc}$	2/3	total strength
	$btob_{sd}$	4	standard deviation
BtoG	$btog_w$	0.125^*	line integral of weights
	$btog_{sdl}$	0.5	relative s.d. on tangential direction
	$btogo_{sds}$	0.25	relative s.d. on radial direction (object)
	$btogc_{sds}$	0.1	relative s.d. on radial direction (contour)
	r	16	radius (input pixels)
IGtoG	$igtog_w$	$-1/8$	line integral of weights
	$igtog_{wo}$	$-1/8$	line integral of weights
GtoB	$gtob_w$	2/3	line integral of weights
GtoIG	$gtoig_w$	1/3	line integral of weights
GtoE	$gtoe_w$	8/3	line integral of weights

A.2 Supplementary figures

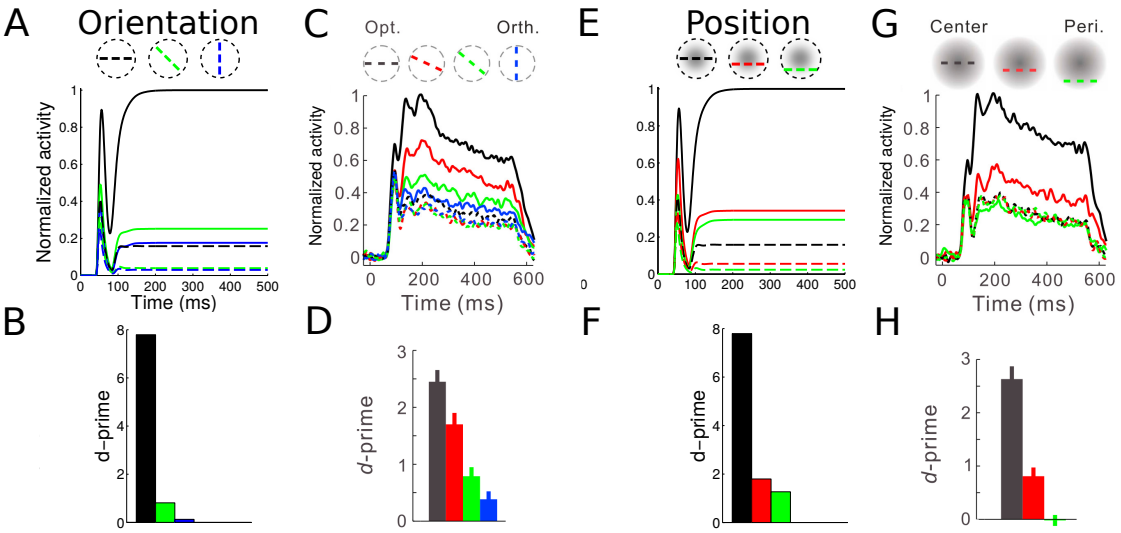


Figure A.1: Orientation and position dependence of contour integration in V4 G_c cells. The top row shows neuronal responses and the bottom row the contour-response d'. Line colors for each figure are indicated by the legends at the top of each column. Top row, solid lines indicate responses for the 7-bar contour pattern, while dashed lines indicate responses for the 1-bar noise pattern. Note that orientations were changed in variable steps based on the tuning curve of the neuron under study in the experimental data (panels C, D) while our simulation only allowed steps of $\pi/4$ (panels A,B). (A and B) Model results, orientation dependence. The neuronal responses (A) and the contour-response d' (B) decreased when the contour was rotated away from the preferred orientation. (C and D) Analogous experimental results, adapted from Chen et al. (2014). (E and F) Model results, position dependence. The neuronal responses (E) and contour-response d' (F) decreased when the contour was translated away from the center of the V4 RF. (G and H) Analogous experimental results, adapted from Chen et al. (2014). Panels C, D, G, and H are modified from Figure 3 of Chen et al. (2014).

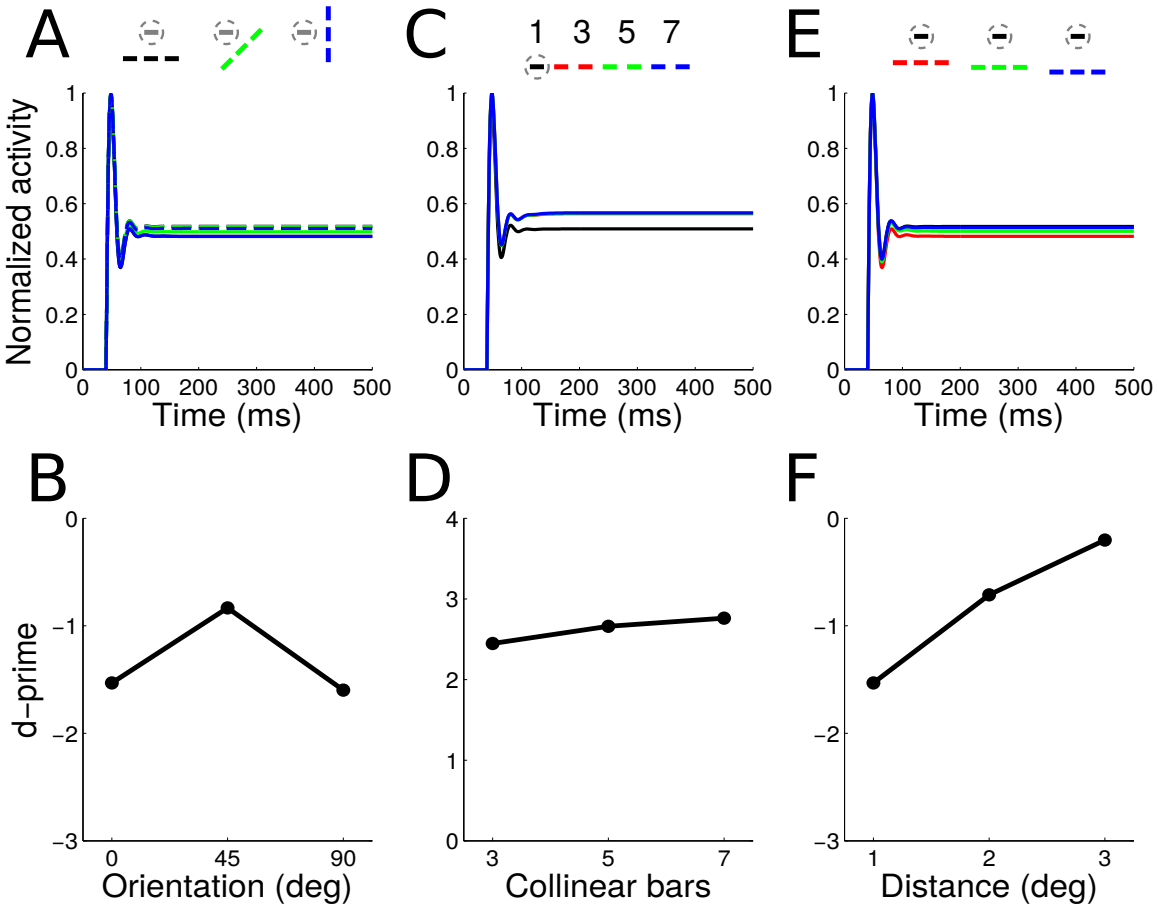


Figure A.2: Orientation and position dependence of contour integration in V1 E cells, model results. The top row shows neuronal responses and the bottom row the contour-response d'. Line colors for each figures are indicated by the legends at the top of each column. (A and B) Orientation dependence of background suppression. The neuronal responses (A) and the contour-response d' (B) increased for intermediate orientations of the background contour. In (A), solid and dashed lines correspond to the 7-bar contour and 1-bar noise patterns, respectively. (C and D) Contour integration on one end. The neuronal responses (C) and contour-response d' (D) increased when bars were added to only one side of the V1 RF. (E and F) Position dependence of background suppression. The neuronal responses (E) and contour-response d' (F) increased (approached zero) when the background contour was moved away from the center of the V1 RF.

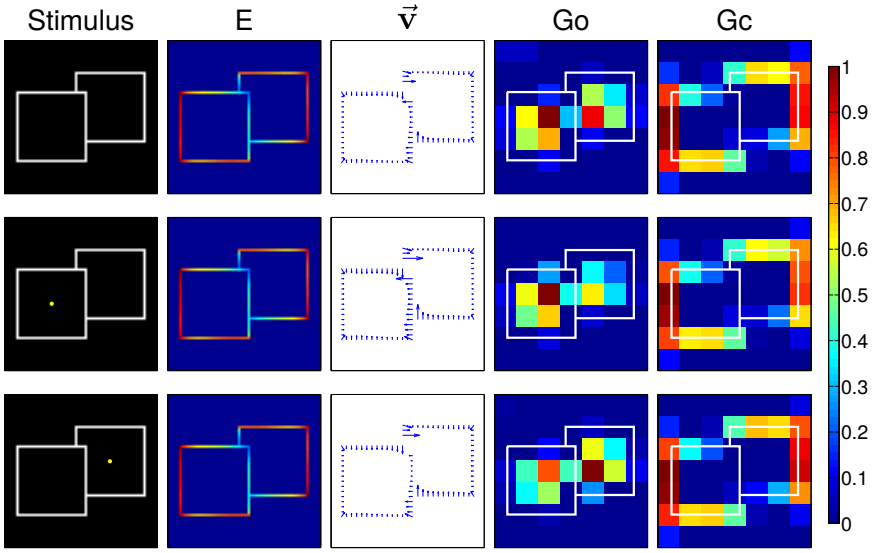


Figure A.3: Attention in the presence of multiple objects. Shown are (left to right) the input stimulus, the edge cell activity (E), the border ownership assignment along edges (shown as the vector modulation index $\vec{\mathbf{v}}$, section 2.2.5), the object grouping neuron activity (Go), and the contour grouping activity (Gc). The input stimulus is overlaid (white lines) on top of the grouping cell activities (columns 4 and 5) to aid visualizing the location of the two figures. Activities are normalized within each map, and warmer colors indicate higher activity (see color bar at right). The yellow dot indicates the locus of attention, which was applied at the level of grouping neurons. In the absence of attention (top row) or when attention is directed to the foreground square (middle row), the edge between the figures is correctly assigned to the foreground. If attention is directed toward the background square (bottom row), the border ownership signal of this edge is greatly reduced, consistent with experimental observations (Qiu et al., 2007).

Even in the absence of attention, figure-ground segregation can be observed in the border-ownership assignment along the edges of the two squares, as well as in the object and contour grouping cell activity (Figure A.3, row 1). Figure A.3 shows the responses of different populations of neurons in our model for the different attention conditions. This figure should be compared with Figure 3 in Mihalas et al. (2011) which is a model related to ours but which doesnot contain G_c cells. The figure

shows that the approximate locations of the foreground and background squares are represented by two peaks in the activity of object grouping neurons (fourth column). Selectively attending to one object enhances the activity of grouping neurons corresponding to the attended object while simultaneously suppressing the activity of grouping neurons representing the unattended object. Attentional modulation at the grouping neuron level is then propagated back to border-ownership selective neurons along object boundaries via the feedback grouping circuitry of our model. Of particular interest is the center edge separating the foreground and background squares. When attention is directed to the foreground square, assignment of border ownership along this edge is strengthened relative to the unattended case (Figure A.3, row 2). However, if attention is directed toward the background square, border-ownership modulation of the edge between the two squares is greatly reduced (Figure A.3, row 3). These results are in agreement with the physiological evidence (Qiu et al., 2007), and demonstrate that grouping mechanisms provide a means to attend to objects in clutter. Functionally, when attention is directed towards the occluded object, suppression of the border-ownership signal along the occluding edge is useful because this edge is "owned" by the occluder and should not be included in the representation of the attended object (Craft et al., 2007).

Bibliography

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):898–916, May 2011. ISSN 0162-8828. doi: 10.1109/TPAMI.2010.161. URL `http://dx.doi.org/10.1109/TPAMI.2010.161`.

D. Ardila, S. Mihalas, R. von der Heydt, and E. Niebur. Medial axis generation in a model of perceptual organization. In *IEEE CISS-2012 46th Annual Conference on Information Sciences and Systems*, pages 1–4, Princeton University, NJ, 2012. IEEE.

George Azzopardi, Antonio Rodríguez-Sánchez, Justus Piater, and Nicolai Petkov. A push-pull corf model of a simple cell with antiphase inhibition improves snr and contour detection. *PloS one*, 9(7):e98424, 2014.

K. Baek and P. Sajda. Inferring figure-ground using a recurrent integrate-and-fire neural circuit. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(2):125–130, 2005.

BIBLIOGRAPHY

Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 1995. URL `http://www.jstor.org/stable/2346101`.

PO Bishop and JD Pettigrew. Neural mechanisms of binocular vision. *Vision research*, 26(9):1587–1600, 1986.

Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

Ali Borji, Dicky N Sihite, and Laurent Itti. Objects do not predict fixations better than early saliency: A re-analysis of Einhäuser et al.'s data. *Journal of vision*, 13 (10):18, 2013.

William H Bosking, Ying Zhang, Brett Schofield, and David Fitzpatrick. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *The Journal of Neuroscience*, 17(6):2112–2127, 1997.

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

Neil DB Bruce and John K Tsotsos. An attentional framework for stereo vision. In *Computer and Robot Vision, 2005. Proceedings. The 2nd Canadian Conference on*, pages 88–95. IEEE, 2005.

M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level

saliency combined with face detection. *Advances in Neural Information Processing Systems*, 20:241–248, 2008.

Garga Chatterjee, Daw-An Wu, and Bhavin R Sheth. Phantom flashes caused by interactions across visual space. *Journal of vision*, 11(2):14, 2011.

Minggui Chen, Yin Yan, Xiajing Gong, Charles D Gilbert, Hualou Liang, and Wu Li. Incremental integration of global contours through interplay between visual cortical areas. *Neuron*, 82(3):682–694, 2014.

Michele A Cox, Michael C Schmid, Andrew J Peters, Richard C Saunders, David A Leopold, and Alexander Maier. Receptive field focus of visual area V4 neurons determines responses to illusory surfaces. *Proceedings of the National Academy of Sciences*, 110(42):17095–17100, 2013.

E. Craft, H. Schütze, E. Niebur, and R. von der Heydt. A neural model of figure-ground organization. *Journal of Neurophysiology*, 97(6):4310–26, 2007. PMID: 17442769.

BG Cumming and GC DeAngelis. The physiology of stereopsis. *Annual review of neuroscience*, 24(1):203–238, 2001. PMID: 11283310.

Sachin S Deshmukh and James J Knierim. Influence of local objects on hippocampal representations: Landmark vectors and memory. *Hippocampus*, 2013. PMID: 23447419 [PubMed - as supplied by publisher].

BIBLIOGRAPHY

James J DiCarlo, Kenneth O Johnson, and Steven S Hsiao. Structure of receptive fields in area 3b of primary somatosensory cortex in the alert monkey. *Journal of neuroscience*, 18(7):2626–2645, 1998.

Piotr Dollár and C Lawrence Zitnick. Fast edge detection using structured forests. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1558–1570, 2015.

Dražen Domijan and Mia Šetić. A feedback model of figure-ground assignment. *Journal of Vision*, 8(7):10–10, 2008.

Y. Dong, S. Mihalas, F. Qiu, R. von der Heydt, and E. Niebur. Synchrony and the binding problem in macaque visual cortex. *Journal of Vision*, 8(7):1–16, 2008. URL `http://journalofvision.org/8/7/30/,doi:10.1167/8.7.30`. PMC2647779.

J. Driver and G.C. Baylis. Edge-assignment and figure-ground segmentation in short-term visual matching. *Cognit Psychol*, 31(3):248–306, 1996.

J. Duncan. Selective attention and the organization of visual information. *J Exp Psychol Gen*, 113:501–517, Dec 1984.

R. Egly, J. Driver, and R. Rafal. Shifting visual attention between objects and locations: evidence for normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, 123:161–77, 1994.

BIBLIOGRAPHY

W. Einhäuser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *J. Vision*, 8(14):1–26, 2008.

Boris Epshtein, Ita Lifshitz, and Shimon Ullman. Image interpretation by a single bottom-up top-down cycle. *Proceedings of the National Academy of Sciences*, 105 (38):14298–14303, 2008.

D. J. Field, A. Hayes, and R. F. Hess. Contour integration by the human visual system: evidence for a local association field. *Vision Research*, 33(2):173–193, 1993.

Charless C Fowlkes, David R Martin, and Jitendra Malik. Local figure–ground cues are valid for natural images. *Journal of Vision*, 7(8):2–2, 2007.

Josselin Gautier and Olivier Le Meur. A time-dependent saliency model combining center and depth biases for 2D and 3D viewing conditions. *Cognitive Computation*, 4(2):141–156, 2012.

Ariel Gilad, Elhanan Meirovithz, and Hamutal Slovin. Population responses to contour integration: early encoding of discrete elements and late perceptual grouping. *Neuron*, 78(2):389–402, 2013.

C.D. Gilbert and T.N. Wiesel. Columnar specificity of intrinsic horizontal and cortico-cortical connections in cat visual cortex. *J. Neurosci.*, 9:2432–2442, 1989.

BIBLIOGRAPHY

P. Girard, J.M. Hupé, and J. Bullier. Feedforward and feedback connections between areas V1 and V2 of the monkey have similar rapid conduction velocities. *J. Neurophysiol.*, 85:1328–1331, 2001.

A Glennerster and SP McKee. Bias and sensitivity of stereo judgements in the presence of a slanted reference plane. *Vision Research*, 39(18):3057–3069, 1999.

D. M. Green and J. A. Swets. *Signal Detection Theory and Psychophysics*. Wiley, New York, N.Y., 1966.

S. Grossberg. 3-D vision and figure-ground separation by visual cortex. *Perception & Psychophysics*, 55:48–120, 1994.

S. Grossberg. Cortical dynamics of three-dimensional figure–ground perception of two-dimensional pictures. *Psychological review*, 104(3):618, 1997. PMID: 9243966.

Z. J. He and K. Nakayama. Surfaces versus features in visual search. *Nature*, 359: 231–233, 1992. PMID: 1528263.

Z. J. He and K. Nakayama. Visual attention to surfaces in three-dimensional space. *Proc. Natl. Acad. Sci. U. S. A.*, 9(24):11155–11159, 1995. PMID: 7479956.

Zijiang J He and Teng Leng Ooi. Perceiving binocular depth with reference to a common surface. *PERCEPTION-LONDON-*, 29(11):1313–1334, 2000.

Jay Hegdé and David C Van Essen. A comparative study of shape representation in macaque visual areas V2 and V4. *Cerebral Cortex*, 17(5):1100–1116, 2007.

BIBLIOGRAPHY

F. Heitger and R. von der Heydt. A computational model of neural contour processing: figure-ground segregation and illusory contours. In *Proc. 4th Int. Conf. Computer Vision*, Proc. 4th Int. Conf. Computer Vision, pages 32–40. IEEE Computer Society Press, 1993.

Janis K Hesse and Doris Y Tsao. Consistency of border-ownership cells across artificial stimuli, natural stimuli, and stimuli with ambiguous contours. *Journal of Neuroscience*, 36(44):11338–11349, 2016.

Ming-Chou Ho and Su-Ling Yeh. Effects of instantaneous object input and past experience on object-based attention. *Acta psychologica*, 132(1):31–39, 2009.

Shaul Hochstein and Merav Ahissar. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5):791–804, 2002.

J.M. Hopf, CN Boehler, SJ Luck, JK Tsotsos, H.J. Heinze, and MA Schoenfeld. Direct neurophysiological evidence for spatial suppression surrounding the focus of attention in vision. *Proceedings of the National Academy of Sciences*, 103(4):1053, 2006. PMID: 16410356.

B. Hu, R. von der Heydt, and E. Niebur. A neural model for perceptual organization of 3D surfaces. In *IEEE CISS-2015 49th Annual Conference on Information Sciences and Systems*, pages 1–6, Baltimore, MD, 2015. IEEE Information Theory Society. doi: 10.1109/CISS.2015.7086906.

BIBLIOGRAPHY

Brian Hu and Ernst Niebur. A recurrent neural model of proto-object based contour integration and figure-ground segregation. *Journal of Computational Neuroscience*, 2017. ISSN 1573-6873. doi: 10.1007/s10827-017-0659-3.

Brian Hu, Ralinkae Kane-Jackson, and Ernst Niebur. A proto-object based saliency model in three-dimensional space. *Vision Research*, 119:42–49, 2016. doi: 10.1016/j.visres.2015.12.004.

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

Heinz Hügli, Timothée Jost, and Nabil Ouerhani. Model performance for visual attention in real 3D color scenes. In *Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach*, pages 469–478. Springer, 2005.

J.M. Hupé, A.C. James, B.R. Payne, S.G. Lomber, P. Girard, and J. Bullier. Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature*, 394:784–787, 1998.

Quan Huynh-Thu and Luca Schiatti. Examination of 3D visual attention in stereoscopic video content. In *IS&T/SPIE Electronic Imaging*, pages 78650J–78650J. International Society for Optics and Photonics, 2011.

J. Intriligator and P. Cavanagh. The spatial resolution of visual attention. *Cognit. Psychol.*, 43:171–216, 2001.

BIBLIOGRAPHY

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

Lina Jansen, Selim Onat, and Peter König. Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision*, 9(1):29, 2009.

Ana Karla Jansen-Amorim, Mario Fiorani, and Ricardo Gattass. GABA inactivation of area V4 changes receptive-field properties of V2 neurons in Cebus monkeys. *Experimental Neurology*, 235(2):553–562, 2012.

Janneke FM Jehee, Victor AF Lamme, and Pieter R Roelfsema. Boundary assignment in a recurrent network architecture. *Vision research*, 47(9):1153–1165, 2007.

Timothée Jost, Nabil Ouerhani, Roman von Wartburg, René Müri, and Heinz Hügli. Contribution of depth to visual attention: comparison of a computer model and human. In *Proceedings. Early cognitive vision workshop*, pages 1–4, 2004.

Tilke Judd, Frédo Durand, and Antonio Torralba. A Benchmark of Computational Models of Saliency to Predict Human Fixations. In *MIT Technical Report*, 2012.

M. Kikuchi and Y. Akashi. A model of border-ownership coding in early vision. In G. Dorffner, H. Bischof, and K. Hornik, editors, *ICANN 2001*, pages 1069–74, 2001.

R. Kimchi, Y. Yeshurun, and A. Cohen-Savransky. Automatic, stimulus-driven attentional capture by objecthood. *Psychon Bull Rev*, 14(1):166–172, Feb 2007.

BIBLIOGRAPHY

Hee-kyoung Ko and Rüdiger von der Heydt. Figure-ground organization in the visual cortex: does meaning matter? *Journal of Neurophysiology*, pages jn–00131, 2017.

C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiol.*, 4:219–227, 1985.

K. Koffka. *Principles of Gestalt psychology.* Harcourt-Brace, New York, 1935.

V. A. F. Lamme. The neurophysiology of figure-ground segregation in primary visual cortex. *J Neurosci*, 15:1605–1615, 1995.

V. A. F. Lamme, K. Zipser, and H. Spekreijse. Figure-ground activity in primary visual cortex is suppressed by anesthesia. *Proc. Natl. Acad. Sci. U. S. A.*, 9(6): 3263–3268, 1998.

Congyan Lang, Tam V Nguyen, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Shuicheng Yan. Depth matters: Influence of depth cues on visual saliency. In *Computer Vision–ECCV 2012*, pages 101–115. Springer, 2012.

Oliver W Layton, Ennio Mingolla, and Arash Yazdanbakhsh. Dynamic coding of border-ownership in visual cortex. *Journal of vision*, 12(13):8, 2012.

P. Le Callet and E. Niebur. Visual Attention and Applications in Multimedia Technologies. *IEEE Proceedings*, 101(9):2058–67, 2013. NIHMS539064.

Tai Sing Lee, David Mumford, Richard Romero, and Victor A. F. Lamme. The Role of

the Primary Visual Cortex in Higher Level Vision. *Vision Research*, 38:2429–2452, 1998.

Sidney R Lehky and Terrence J Sejnowski. Neural model of stereoacuity and depth interpolation based on a distributed representation of stereo disparity. *Journal of Neuroscience*, 10(7):2281–2299, 1990.

Ido Leichter and Michael Lindenbaum. Boundary ownership by lifting to 2.1 d. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 9–16. IEEE, 2009.

Michael S Lewicki, Bruno A Olshausen, Annemarie Surlykke, and Cynthia F Moss. Scene analysis in the natural environment. *Frontiers in psychology*, 5, 2014.

Wu Li, Valentin Piëch, and Charles D Gilbert. Perceptual learning and top-down influences in primary visual cortex. *Nature neuroscience*, 7(6):651–657, 2004.

Wu Li, Valentin Piëch, and Charles D Gilbert. Contour saliency in primary visual cortex. *Neuron*, 50(6):951–962, 2006.

Wu Li, Valentin Piëch, and Charles D Gilbert. Learning to link visual contours. *Neuron*, 57(3):442–451, 2008.

Z. Li. A Neural Model of Contour Integration in the Primary Visual Cortex. *Neural Computation*, 10(903-940), 1998.

BIBLIOGRAPHY

Chih-Yao Ma and Hsueh-Ming Hang. Learning-based saliency model with depth information. *Journal of vision*, 15(6):19–19, 2015.

Joseph R Manns and Howard Eichenbaum. A cognitive map for object memory in the hippocampus. *Learning & Memory*, 16(10):616–624, 2009. PMC2769165.

D. Marr and T. Poggio. Cooperative Computation of Stereo Disparity. *Science*, 194, 1976. PMID: 968482.

Jonathan A Marshall, George J Kalarickal, and Elizabeth B Graves. Neural model of visual stereomatching: slant, transparency and clouds. *Network: Computation in Neural Systems*, 7(4):635–669, 1996.

Anne B Martin and Rüdiger von der Heydt. Spike Synchrony Reveals Emergence of Proto-Objects in Visual Cortex. *The Journal of Neuroscience*, 35(17):6860–6870, 2015.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.

C. J. McAdams and J. H. R. Maunsell. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J. Neurosci.*, 19:431–441, 1999.

BIBLIOGRAPHY

Jason S McCarley and Zijiang J He. Sequential priming of 3-d perceptual organization. *Attention, Perception, & Psychophysics*, 63(2):195–208, 2001.

Suzanne P McKee. The spatial requirements for fine stereoacuity. *Vision research*, 23(2):191–198, 1983.

S. Mihalas, Y. Dong, R. von der Heydt, and E. Niebur. Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects. *Proceedings of the National Academy of Sciences*, 108(18):7583–8, 2011. PMC3088583.

Ajay K Mishra, Ashish Shrivastava, and Yiannis Aloimonos. Segmenting "simple"??? objects using RGB-D. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 4406–4413. IEEE, 2012.

B. C. Motter. Neural correlates of attentive selection for color or luminance in extrastriate area V4. *J. Neurosci.*, 14:2178–2189, 1994.

B.C. Motter. Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *J. Neurophysiology*, 70 (3):909–919, 1993.

K. Nakayama and G. H. Silverman. Serial and parallel processing of visual feature conjunctions. *Nature*, 320:264–265, 1986. PMID: 3960106.

K. Nakayama, Z. J. He, and S. Shimojo. Visual surface representation: a critical link between lower-level and higher-level vision. In S. Kosslyn and D. Osherson, editors,

Visual Cognition: An Invitation to Cognitive Science, volume 2, chapter 1, pages 1–70. The MIT Press, 2nd edition, 1995.

E. Niebur. Separate but equal: Different kinds of information require different neural representations. In H. Bothe, editor, *Proceedings of the International Congress on Intelligent Systems and Applications (ISA-BIS)*, pages 1544–9, Wetaskiwin, Canada, December 2000. ICSC Academic Press. ISBN 3-906454-24-X.

E. Niebur and C. Koch. Control of Selective Visual Attention: Modeling the "Where" Pathway. In D. S Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 802–808. MIT Press, Cambridge, MA, 1996.

H. Nishimura and K. Sakai. Determination of border-ownership based on the surround context of contrast. *Neurocomputing*, 58-60:843–8, 2004.

H. Nishimura and K. Sakai. The computational model for border-ownership determination consisting of surrounding suppression and facilitation in early vision. *Neurocomputing*, 65:77–83, 2005.

Antje Nuthmann and John M Henderson. Object-based attentional selection in scene viewing. *Journal of vision*, 10(8):20, 2010.

P.J. O'Herron and R. von der Heydt. Short-term memory for figure-ground organization in the visual cortex. *Neuron*, 61(5):801–809, 2009. PMC2707495.

BIBLIOGRAPHY

P.J. O'Herron and R. von der Heydt. Remapping of Border Ownership in the Visual Cortex. *The Journal of Neuroscience*, 33(5):1964–1974, 2013. PMID: 23365235 [PubMed - in process].

I. Ohzawa, G. C. DeAngelis, and R. D. Freeman. Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. *Science*, 249: 1037–1041, 1990.

Teng Leng Ooi and Zijiang J He. Space perception of strabismic observers in the real world environmentspace perception of strabismic observers. *Investigative ophthalmology & visual science*, 56(3):1761–1768, 2015.

Nabil Ouerhani and Heinz Hügli. Computing visual attention from scene depth. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 1, pages 375–378. IEEE, 2000.

Lucy M Palmer, Adam S Shai, James E Reeve, Harry L Anderson, Ole Paulsen, and Matthew E Larkum. NMDA spikes enhance action potential generation during sensory input. *Nature neuroscience*, 17(3):383–390, 2014.

S.E. Palmer and T. Ghose. Extremal Edge– A Powerful Cue to Depth Perception and Figure-Ground Organization. *Psychological Science*, 19(1):77, 2008. ISSN 0956-7976.

BIBLIOGRAPHY

Stephen E Palmer. Perceptual organization in vision. *Stevens' handbook of experimental psychology*, 2002.

H.-K. Pao, D. Geiger, and N Rubin. Measuring convexity for Figure/Ground Separation. In *7th International Conference on Computer Vision*, Kerkyra, Greece, September 1999.

D. Parkhurst, K. Law, and E. Niebur. Modelling the role of salience in the allocation of visual selective attention. *Vision Research*, 42(1):107–123, 2002.

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

Valentin Piëch, Wu Li, George N Reeke, and Charles D Gilbert. Network model of top-down influences on local gain and contextual interactions in visual cortex. *Proceedings of the National Academy of Sciences*, 110(43):E4108–E4117, 2013. PMC3808648.

G. F. Poggio and B. Fischer. Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey. *J. Neurophysiol.*, 40:1392–1405, Nov 1977. PMID: 411898.

G.F. Poggio and T. Poggio. The analysis of stereopsis. *Ann. Rev. Neurosci.*, 7: 379–412, 1984.

BIBLIOGRAPHY

Gian F Poggio, Francisco Gonzalez, and F Krause. Stereoscopic mechanisms in monkey visual cortex: binocular correlation and disparity selectivity. *The Journal of neuroscience*, 8(12):4531–4550, 1988.

Uri Polat, Keiko Mizobe, Mark W. Pettet, Takuji Kasamatsu, and Anthony M. Norcia. Collinear Stimuli Regulate Visual Responses Depending on Cell's Contrast Threshold. *Nature*, 391:580–584, February 1998.

Jasper Poort, Florian Raudies, Aurel Wannig, Victor AF Lamme, Heiko Neumann, and Pieter R Roelfsema. The role of attention in figure-ground segregation in areas V1 and V4 of the visual cortex. *Neuron*, 75(1):143–156, 2012.

Alexandre Pouget and Terrence J. Sejnowski. Spatial Transformations in the Parietal Cortex Using Basis Functions. *Journal of Cognitive Neuroscience*, 9(2):222–237, 1997. PMID: 23962013.

F. T. Qiu and R. von der Heydt. Figure and ground in the visual cortex: V2 combines stereoscopic cues with Gestalt rules. *Neuron*, 47:155–166, 2005.

F. T. Qiu and R. von der Heydt. Neural representation of transparent overlay. *Nat. Neurosci.*, 10(3):283–284, 2007.

F. T. Qiu, T. Sugihara, and R. von der Heydt. Understanding the neural mechanisms of object-based visual attention. *Soc. Neurosci. Abstr.*, page 821.13, 2005.

BIBLIOGRAPHY

F. T. Qiu, T. Sugihara, and R. von der Heydt. Figure-ground mechanisms provide structure for selective attention. *Nat. Neurosci.*, 10(11):1492–9, October 2007.

S. Ramenahalli and E. Niebur. Computing 3D saliency from a 2D image. In *Information Sciences and Systems (CISS), 2013 47th Annual Conference on*, pages 1–5, 2013. doi: 10.1109/CISS.2013.6552297. URL `http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6552297`.

S. Ramenahalli, S. Mihalas, and E. Niebur. Extremal Edges: Evidence in Natural Images. In *IEEE CISS-2011 45th Annual Conference on Information Sciences and Systems*, pages 1–6, Baltimore, MD, 2011. IEEE Information Theory Society.

S. Ramenahalli, S. Mihalas, and E. Niebur. Figure-ground classification based on spectral properties of boundary image patches. In *IEEE CISS-2012 46th Annual Conference on Information Sciences and Systems*, pages 1–6, Princeton, NJ, 2012. IEEE Information Theory Society.

Sudarshan Ramenahalli, Stefan Mihalas, and Ernst Niebur. Local spectral anisotropy is a valid cue for figure-ground organization in natural scenes. *Vision Research*, 103: 116–126, Oct 2014. doi: 10.1016/j.visres.2014.08.012. URL `http://dx.doi.org/10.1016/j.visres.2014.08.012`. NIHMSID 631573.

Xiaofeng Ren, Charless C Fowlkes, and Jitendra Malik. Figure/ground assignment in natural images. In *Proceedings of the 9th European conference on Computer Vision-Volume Part II*, pages 614–627. Springer-Verlag, 2006.

BIBLIOGRAPHY

R. A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7(1/2/3): 17–42, 2000.

Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Saliency and human fixations: state-of-the-art and study of comparison metrics. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1153–1160. IEEE, 2013.

P. R. Roelfsema, V. A. F. Lamme, and H. Spekreijse. Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, 395(6700):376–381, 1998.

Pieter R Roelfsema. Cortical algorithms for perceptual grouping. *Annu. Rev. Neurosci.*, 29:203–227, 2006.

Pieter R Roelfsema, Victor A F Lamme, and Henk Spekreijse. Synchrony and covariation of firing rates in the primary visual cortex during contour grouping. *Nat Neurosci*, 7(9):982–991, Sep 2004. doi: 10.1038/nn1304. URL `http://dx.doi.org/10.1038/nn1304`.

Ari Rosenberg, Noah J Cowan, and Dora E Angelaki. The Visual Representation of 3D Object Orientation in Parietal Cortex. *The Journal of Neuroscience*, 33(49): 19352–19361, 2013.

A. F. Russell, S Mihalas, R. von der Heydt, E. Niebur, and R. Etienne-Cummings. A model of proto-object based saliency. *Vision Research*, 94:1–15, 2014.

BIBLIOGRAPHY

P Sajda and L.H. Finkel. Intermediate-Level Visual Representations and the Construction of Surface Perception. *J Cogn Neurosci*, 7:267–291, 1995.

K. Sakai and H. Nishimura. Surrounding suppression and facilitation in the determination of border ownership. *Journal of Cognitive Neuroscience*, 18(4):562–579, 2006.

Ko Sakai, Haruka Nishimura, Ryohei Shimizu, and Keiichi Kondo. Consistent and robust determination of border ownership based on asymmetric surrounding contrast. *Neural Networks*, 33:257–274, 2012.

Emilio Salinas and L. F. Abbott. Transfer of coded information from sensory to motor networks. *J. Neurosci.*, 15(10):6461–6474, October 1995. PMID: 7472409.

Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(5):824–840, 2009.

Daniel Scharstein and Chris Pal. Learning conditional random fields for stereo. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

Mircea A Schoenfeld, Jens-Max Hopf, Christian Merkel, Hans-Jochen Heinze, and Steven A Hillyard. Object-based attention involves the sequential activation of feature-specific cortical modules. *Nature neuroscience*, 17(4):619–624, 2014.

BIBLIOGRAPHY

B. J. Scholl. Objects and attention: the state of the art. *Cognition*, 80(1-2):1–46, 2001.

H Schütze, E. Niebur, and R. von der Heydt. Modeling cortical mechanisms of border ownership coding. *J. Vision*, 3(9):114a, 2003.

M. Sigman, G. A. Cecchi, C. D. Gilbert, and M. O. Magnasco. On a common circle: natural scenes and Gestalt rules. *Proc Natl Acad Sci U S A*, 98(4):1935–1940, Feb 2001. doi: 10.1073/pnas.031571498. URL `http://dx.doi.org/10.1073/pnas.031571498`.

Enrico Simonotto, Massimo Riani, Charles Seife, Mark Roberts, Jennifer Twitty, and Frank Moss. Visual perception of stochastic resonance. *Physical Review Letters*, 78(6):1186, 1997.

W. Singer. Neuronal synchrony: a versatile code for the definition of relations? *Neuron*, 24:49–65, 1999.

M. Stemmler, M. Usher, and E. Niebur. Lateral cortical connections may contribute to both contour completion and redundancy reduction in visual processing. *Soc. Neurosci. Abstr.*, 21(1):510, 1995a.

M. Stemmler, M. Usher, and E. Niebur. Lateral Interactions in Primary Visual Cortex: A Model Bridging Physiology and Psychophysics. *Science*, 269:1877–1880, 1995b.

BIBLIOGRAPHY

Dan D Stettler, Aniruddha Das, Jean Bennett, and Charles D Gilbert. Lateral connectivity and contextual interactions in macaque primary visual cortex. *Neuron*, 36(4):739–750, 2002.

Scott B Stevenson, Lawrence K Cormack, and Clifton M Schor. Depth attraction and repulsion in random dot stereograms. *Vision research*, 31(5):805–813, 1991.

Josef Stoll, Michael Thrun, Antje Nuthmann, and Wolfgang Einhäuser. Overt attention in natural scenes: objects dominate features. *Vision research*, 107:36–48, 2015.

Tadashi Sugihara, Fangtu T Qiu, and Rüdiger von der Heydt. The speed of context integration in the visual cortex. *Journal of neurophysiology*, 106(1):374–385, 2011. PMC3129740.

K. A. Sundberg, J. F. Mitchell, and J. H. Reynolds. Spatial attention modulates center-surround interactions in macaque visual area v4. *Neuron*, 61:952–963, Mar 2009.

Hans Supèr and Victor AF Lamme. Altered figure-ground perception in monkeys with an extra-striate lesion. *Neuropsychologia*, 45(14):3329–3334, 2007.

Hans Supèr, August Romeo, and Matthias Keil. Feed-forward segmentation of figure-ground and assignment of border-ownership. *PLoS One*, 5(5):e10705, 2010.

BIBLIOGRAPHY

Ching Teo, Cornelia Fermuller, and Yiannis Aloimonos. Fast 2d border ownership assignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5117–5125, 2015.

A. Thiele and G. Stoner. Neuronal synchrony does not correlate with motion coherence in cortical area MT. *Nature*, 421(6921):366–370, 2003.

A. Treisman. The binding problem. *Curr Opin Neurobiol*, 6(2):171–178, April 1996.

A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980. PMID: 7351125.

J. C. M. Treue, S.and Trujillo. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399:575–9, 1999.

S. Treue. Neural correlates of attention in primate visual cortex. *Trends in Neurosciences*, 24:295–300, May 2001.

Stephan Tschechne and Heiko Neumann. Hierarchical representation of shapes in visual cortexfrom localized features to figural shape segregation. *Frontiers in computational neuroscience*, 8:93, 2014.

Inna Tsirlin, Robert S Allison, and Laurie M Wilcox. Stereoscopic transparency: Constraints on the perception of multiple surfaces. *Journal of Vision*, 8(5), 2008.

J. K. Tsotsos. Analyzing vision at the complexity level. *Behav. Brain Sci.*, 13: 423–469, 1990.

BIBLIOGRAPHY

John K Tsotsos. *A Computational Perspective on Visual Attention*. MIT Press, 2011.

S. Ullman. Visual routines. *Cognition*, 18:97–159, 1984.

Shimon Ullman, RL Gregory, and J Atkinson. Low-Level Aspects of Segmentation and Recognition [and Discussion]. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 337(1281):371–379, 1992.

Leslie G Ungerleider, Thelma W Galkin, Robert Desimone, and Ricardo Gattass. Cortical connections of area V4 in the macaque. *Cerebral Cortex*, 18(3):477–499, 2007.

Matthieu Urvoy, Marcus Barkowsky, Romain Cousseau, Yao Koudota, Vincent Ricorde, Patrick Le Callet, Jesus Gutierrez, and Narciso Garcia. NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences. In *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, pages 109–114. IEEE, 2012.

Shaun P Vecera, Anastasia V Flevaris, and Joseph C Filapek. Exogenous spatial attention influences figure-ground assignment. *Psychological Science*, 15(1):20–26, 2004.

R. von der Heydt, H. Zhou, and H. S. Friedman. Representation of stereoscopic edges in monkey visual cortex. *Vision Res.*, 40(15):1955–1967, 2000. PMID: 10828464.

BIBLIOGRAPHY

R. von der Heydt, F. T. Qiu, and Z. J. He. Neural mechanisms in border ownership assignment: motion parallax and gestalt cues. *J. Vision*, 3(9):666a, 2003.

N. Wagatsuma, R. von der Heydt, and E. Niebur. Spike Synchrony Generated by Modulatory Common Input through NMDA-type Synapses. *Journal of Neurophysiology*, 116(3):1418–1433, 2016.

D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19:1395–1407, Nov 2006.

Junle Wang, Matthieu Perreira DaSilva, Patrick LeCallet, and Vincent Ricordel. Computational model of stereoscopic 3D visual saliency. *Image Processing, IEEE Transactions on*, 22(6):2151–2165, 2013.

Peng Wang and Alan Yuille. Doc: Deep occlusion estimation from a single image. In *European Conference on Computer Vision*, pages 545–561. Springer, 2016.

Detlef Wegener, Winrich A Freiwald, and Andreas K Kreiter. The influence of sustained selective attention on stimulus selectivity in macaque visual area mt. *Journal of Neuroscience*, 24(27):6106–6114, 2004.

Manuel Werlberger, Werner Trobin, Thomas Pock, Andreas Wedel, Daniel Cremers, and Horst Bischof. Anisotropic Huber-L1 Optical Flow. In *Proceedings of the British Machine Vision Conference (BMVC)*, London, UK, September 2009.

BIBLIOGRAPHY

M. Wertheimer. Untersuchungen zur Lehre von der Gestalt II. *Psychol. Forsch.*, 4: 301–350, 1923.

B. Widrow and M.E. Hoff. Adaptive Switching Circuits. 4:96–104, 1960.

Jonathan R Williford and Rudiger von der Heydt. Border-ownership coding. *Scholarpedia*, 8(10):30040, 2013.

Jonathan R Williford and Rudiger von der Heydt. Early Visual Cortex Assigns Border Ownership in Natural Scenes According to Image Context. *Journal of Vision*, 14 (10):588–588, 2014.

Jonathan R Williford and Rüdiger von der Heydt. Figure-ground organization in visual cortex for natural scenes. *eNeuro*, 3(6):ENEURO–0127, 2016.

Jonathan R. Williford and Rdiger von der Heydt. *Data associated with publication Figure-ground organization in visual cortex for natural scenes.*. Johns Hopkins University Data Archive Dataverse, 2017. doi: 10.7281/T1C8276W. URL `http://dx.doi.org/10.7281/T1C8276W`.

Jun Xie, Guanghua Xu, Jing Wang, Sicong Zhang, Feng Zhang, Yeping Li, Chengcheng Han, and Lili Li. Addition of visual noise boosts evoked potential-based brain-computer interface. *Scientific reports*, 4, 2014.

Yin Yan, Malte J Rasch, Minggui Chen, Xiaoping Xiang, Min Huang, Si Wu, and

Wu Li. Perceptual training continuously refines neuronal population codes in primary visual cortex. *Nature neuroscience*, 17(10):1380–1387, 2014.

Shih-Cheng Yen and Leif H Finkel. Extraction of perceptually salient contours by striate cortical networks. *Vision research*, 38(5):719–741, 1998.

Semir M Zeki. Uniformity and diversity of structure and function in rhesus monkey prestriate visual cortex. *The Journal of Physiology*, 277(1):273–290, 1978.

N.R. Zhang and R. von der Heydt. Analysis of the context integration mechanisms underlying figure–ground organization in the visual cortex. *The Journal of Neuroscience*, 30(19):6482–6496, 2010. PMC2910339.

Li Zhaoping. Border ownership from intracortical interactions in visual area V2. *Neuron*, 47:143–153, 2005. PMID: 15996554.

H. Zhou, H. S. Friedman, and R. von der Heydt. Coding of border ownership in monkey visual cortex. *J. Neurosci.*, 20(17):6594–6611, 2000. PMID: 10964965.

Steven W Zucker. Local field potentials and border ownership: a conjecture about computation in visual cortex. *Journal of Physiology-Paris*, 106(5):297–315, 2012.

Timm Zwickel, Thomas Wachtler, and Reinhard Eckhorn. Coding the presence of visual objects in a recurrent neural network of visual cortex. *Biosystems*, 89(1): 216–226, 2007.