# Limitations of using bags of complex features: Hierarchical higher-order filters fail to capture spatial configurations

Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Master of Arts in the Graduate School of the Ohio State University

By

Nicholas M. Van Horn, B.A.

Graduate Program in Psychology

The Ohio State University

2011

Thesis Committee:

Alexander A. Petrov, Advisor

James T. Todd

Dirk Bernhardt-Walther

## Abstract

One common method of representing images is to reduce an image to a collection of features. Many simple features have been proposed, such as pixel intensities and wavelet responses, but these choices are fundamentally unsuitable for capturing the configural relations of objects and object parts, as spatial information associated with each feature is lost. Another recent strategy, known as "feature-hierarchy" modeling, involves the use of overlapping, redundant features. These features are obtained by processing an image across a hierarchy of units tuned to progressively more complex properties. An open question is whether such approaches produce data structures rich enough for implicitly capturing configural relations. We implemented three experiments and several computer simulations to address this issue. Our method involved the use of four classes of objects, each derived from the simple spatial relationships present in classic Vernier and bisection acuity tasks. All human observers achieved near perfect categorization performance after relatively few exposures to each stimulus class. This ability also transferred across several dimensions, including orientation and background context. By contrast, simulations on a feature-hierarchy model revealed poor performance for this class of models. Furthermore, the moderate categorization accuracy achieved did not transfer across even the simplest of dimensions. These results indicate that this approach to image representation lacks a fundamental property necessary for encoding the spatial configurations of object parts.

*To my wife*

## Acknowledgments

I would like to extend a great deal of thanks to my advisor Alexander Petrov for his guidance before and during this project. Without his advice and support this work in its present form would have never been possible. Also, I owe the origin of this research to James Todd, whose insight and healthy skepticism led me down the rabbit hole.

# Vita

| | |
|---|---|
| 1997 ..... | Clearfork High School |
| 2007 ..... | B.A. Japanese, The Ohio State University |
| 2008-Present ..... | Graduate Teaching Associate, Department of Psychology, The Ohio State University |

## Publications

Van Horn, N. & Petrov, A. (2009). Perceptual learning of visual motion: The role of the spatial frequency of the carrier. *Journal of Vision, 9(8),* 886

Van Horn, N. & Petrov, A. (2009). Motion aftereffect duration is not changed by perceptual learning: Evidence against the representation-modification hypothesis. *Annual meeting of the Psychonomic Society,* 5094

Petrov, A. A. & Van Horn, N. M. & Ratcliff, R. (2011). Dissociable perceptual learning mechanisms revealed by diffusion model analysis. *Psychonomic Bulletin & Review, 18*(3), 490–497 .

Petrov, A. A. & Van Horn, N. M. (under review). Motion aftereffect duration is not changed by perceptual learning: Evidence against the representation-modification hypothesis. *Vision Research.*

## Fields of Study

Major Field: Psychology

# Table of Contents

# List of Tables

# List of Figures

# CHAPTER 1

# Introduction

With advances in computing over the last several decades, computers have begun to surpass humans on many tasks. Despite the overwhelming number of computations that can be completed per second, a human child is capable of vastly outperforming even the best computer at visual object recognition tasks. Because of this, many models of object recognition have begun to look to the human visual system for inspiration (e.g., Fukushima, 1980; Perret & Oram, 1993; Amit & Mascaro, 2003; Wersing & Koerner, 2003). Computationally, the difficultly in deriving an algorithm of object detection and recognition lies in the tradeoff between object selectivity and invariance. This means that for a model to be successful in a variety of applications, it must be able to reliably detect differences between object classes while remaining largely indifferent to individual differences in objects within the same class (for an exception, see DiCarlo & Cox, 2007). This includes variations due to position, scale, orientation, occlusion, illumination, within-class heterogeneity, and more.

Early approaches to object recognition relied on template matching. Under this strategy, every possible object is represented in memory by one or more exemplar templates. Potential objects in an input image are compared to templates in memory and successful matches trigger a response from the model. This approach remained unsuitable for practical applications, as template-matching is brittle with respect to the variability contained in most typical images.

At a basic level, it became clear that a successful model of vision would require the use of basic image properties that are invariant to the types of object variety found in the world. One common strategy for image analysis is to reduce an image to a histogram of "features." The types of features that have been used include individual pixel intensities (Swain & Ballard, 1991), receptive fields (Schiele & Crowley, 1996), local scale invariant features (Lowe, 1999), three-dimensional textons (Leung & Malik, 2001), oriented gradients (Dalal & Triggs, 2005), and more. The basic methodology of these approaches is to pass an image through a bank of filters, effectively stripping the image down to a collection of individual elements. The features are then analyzed and discriminated by the use of a secondary classifier. Such approaches have led to successes on very specialized tasks. However, in the best cases these modeling approaches retain minimalistic local spatial information nearby to extracted features, and in the worst cases, all spatial information is lost. In these extreme cases, images are reduced to a bag of free-floating features. From the perspective of these models, a given feature could have originated from any possible location in the image. In contrived situations this is not a problem, as a large enough collection of disparate features will constrain the problem space enough to allow for accurate classification. Not surprisingly, however, when the spatial arrangement of image features becomes critical to solving a task, these data structures do not provide the necessary information. Such bag-of-feature approaches shift the responsibility of extracting relational information from early visual areas to later visual areas, implying that our awareness of the arrangement of objects or object-parts in a scene is due to higher-level reasoning, rather than a direct perception.

However, there is growing evidence that people do directly perceive spatial configurations in images (e.g., Biederman, 1981, 1987; Logan, 1994; Pylyshyn, 2007; J. Kim & Biederman, 2010). These relations are not just passively encoded either,

but appear to be fundamental to many perceptions. For example, Green & Hummel, 2004 found that when neighboring objects are arranged in a functionally meaningful way (such as a hammer turned toward the head of a nail), object identification is facilitated. However, when objects share similar contextual and spatial relations, but are not functionally arranged (such as a teapot pouring away from a cup), no facilitation occurs. Green & Hummel, 2006 took this idea further, demonstrating that by changing the stimulus onset asynchrony (SOA) between the two objects from 100 msec to 250 msec, facilitation once again is lost. If grouping of objects and object parts occurs downstream during higher-level reasoning, then we would expect the subtle change in SOA to have no effect on identification. The fact that a larger SOA produces this effect suggests that near-simultaneously presented features are bound together into a single cohesive item early in the visual system.

Structural description theories (Humphreys, 1987; Biederman, 1987; Hummel & Biederman, 1992) address this issue by explicitly encoding the spatial information of object parts. Thus, both the individual features and their relative positions are encoded separately. Importantly, both forms of information are bound together, preserving both configural relations and individual features. This approach requires a neural mechanism (von der Malsburg, 1999) capable of simultaneously binding features from the same object into a group, while maintaining inter-object segregation. While the exact mechanism capable of achieving this task remains unknown, recent fMRI data at least indirectly suggests that such binding may be occurring (Bar et al., 2001). In general, the lateral occipital complex has been implicated in the perception of the final representational form of visual objects (J. G. Kim, Biederman, Lescroart, & Hayworth, 2009).

Work by Hayworth, Lescroart, & Biederman, 2010 found evidence that the posterior fusiform (pFs), the most anterior region of the LOC, is actively involved in the

representation of configural relations. A series of fMRI adaptation studies revealed that changes in the relative positions of objects, as opposed to changes in absolute positions of the objects, led to increased BOLD responses. The results are in accord with one possible account that relies on "object files" (Kahneman, Treisman, & Gibbs, 1992). Such approaches, such as the FINST theory of Pylyshyn, 2007, 2009, posit that individual objects (or object parts) are loaded into neural "slots" that collectively capture and manage the relational properties of multiple elements.

While bag-of-feature models may not explicitly encode for configural relations, a particularly popular and successful variation on this class of models may implicitly capture the necessary spatial information. This class of models, known as "feature hierarchy" models, uses dense, overlapping receptive fields to translate an image into an internal representation (Figure 1.1). By doing so, features are redundantly encoded

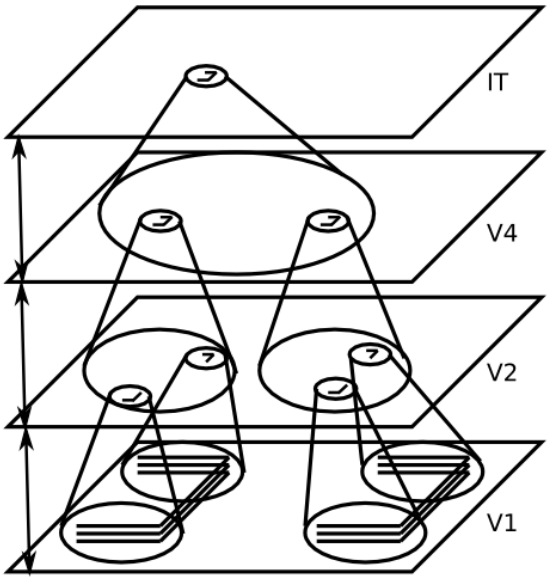


Figure 1.1: Feature hierarchy model schematic. Each level processes increasingly more complex features, possibly allowing for the binding of features and spatial relationships (O'Reilly & Munakata, 2000).

for, giving rise to higher-order complex feature templates which enable an object to be reconstructed downstream. Whereas simple histograms of features throw away all spatial information, one hypothesis is that this redundant coding scheme provides sufficient information for a classification task wherein the configuration of features becomes critical to success. There is an open question as to whether a hierarchical bag-of-features approach to image analysis is sufficient in this way. This was the motivating question for the present study.

To answer this question, we endeavored to create a set of object classes that both minimized feature complexity and relied on spatial configurations of the parts for identification. One reasonable simplification involved two very well understood visual tasks: Vernier and bisection acuity tasks. For Vernier stimuli, an observer is typically asked to resolve progressively smaller displacements in the relative spatial position of disjoint dot patterns or line segments (i.e., whether or not the segments or dots are collinear). Depending on the specific stimuli, observers exhibit thresholds as low as 5 arcsec (Westheimer, 1981; Klein & Levi, 1985). The diameter of the photoreceptors in the retina are around 30–60 arcsec, which means that human observers readily demonstrate hyperacuity at levels less than 1/5 the size of a single photoreceptor. The presence of hyperacuity implies that the spatial relations are being extracted beyond the retinal level (Waugh & Levi, 1995). Similar results have been found for a variety of bisection tasks (Westheimer, Crist, Gorski, & Gilbert, 2001), wherein the goal is to detect if a flanked visual element is equidistant from the two end points, or if it is shifted closer to one end.

Important to the discussion here is not the level of hyperacuity possible, but rather the consistent and strong finding that human observers are able to very accurately perceive the relative spatial positions of separate elements in the visual field. Because hyperacuity on these tasks is such a robust phenomenon, we should expect that any

sufficient model of object recognition should have a mechanism capable of encoding such spatial information. If feature hierarchy-based models are able to capture spatial configurations, then the simple and reliable relations found in visual acuity stimuli should be a given. One reasonable sanity check could involve creating an object classification task that relies on such configurations. With this idea in mind, we implemented several behavioral experiments featuring four object classes built upon these principles. The results of our human performers are then compared to performance from a successful feature hierarchy model known as the "standard," or HMAX model (Riesenhuber & Poggio, 1999).

# CHAPTER 2

# Experiments

## 2.1 Experiment 1

### 2.1.1 Participants

Twelve individuals from The Ohio State University participated for course credit. All subjects were verified to have normal or corrected-to-normal vision by means of a Snellen eye chart.

### 2.1.2 Stimuli and Procedure

Stimuli consisted of four novel object categories, each derived from classic Vernier and bisection tasks (e.g., Seitz et al., 2005; Fahle & Morgan, 1996; Poggio, Fahle, & Edelman, 1992). All four categories were composed of three circular white dots embedded within one of two background contexts (Figure 2.1). Two independent binary factors jointly determined class membership: (1) collinearity of the dots (Vernier), and (2) spacing between dots (bisection). Class 1 objects had collinear dots with equidistant spacing. Class 2 objects featured collinear, but unevenly-spaced dots. Class 3 objects had non-collinear, equally-spaced dots. Class 4 objects had non-collinear, unequally-spaced dots. For categories 2–4, the direction of shift in the center dot relative to the outer dots (i.e., the "bisection" and "Vernier" displacements) could occur in either of two directions. This means, for example, that in the case of stimuli from class 4,

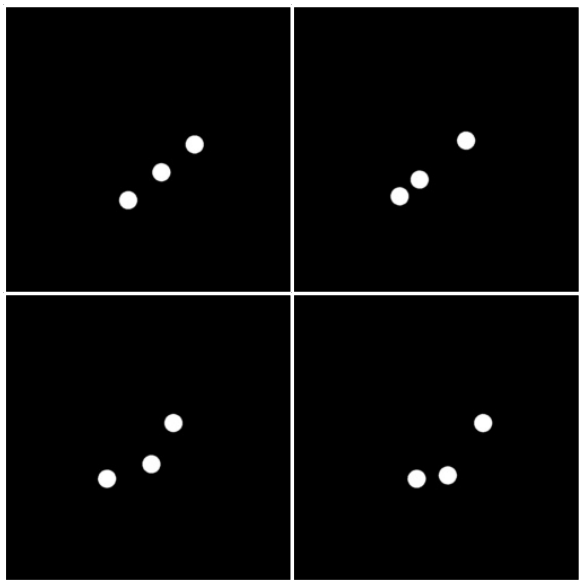

Figure 2.1: Example training stimuli from each of four categories. Subjects trained on stimuli oriented with an implicit angle (here +40° from vertical). Reading across rows: categories 1–4. Actual stimuli varied with respect to the overal scale and position of the dots within the image.

the center dot could be in one of four positions relative to the outer two dots. The scale of the dots varied (as a group) from image to image, and the dots were free to translate within the background context.

Stimuli were generated using Matlab (The MathWorks, 2009) and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997), and were presented at 96 Hz on a gamma-corrected 21" NEC AccuSync 120 color CRT. The experiment was conducted in a dark room, and gaze distance was held constant at 93 cm by means of a chin rest.

The procedure consisted of a 6-block training period followed by a 10-block test period. All blocks contained eight randomly generated exemplars from each class, for a total of 32 total trials per block. During training blocks, the three dots fell along an implicit major axis that was held at a fixed angle across all trials. Half of the subjects trained on an angle −50° from vertical, and half trained on +40°. No

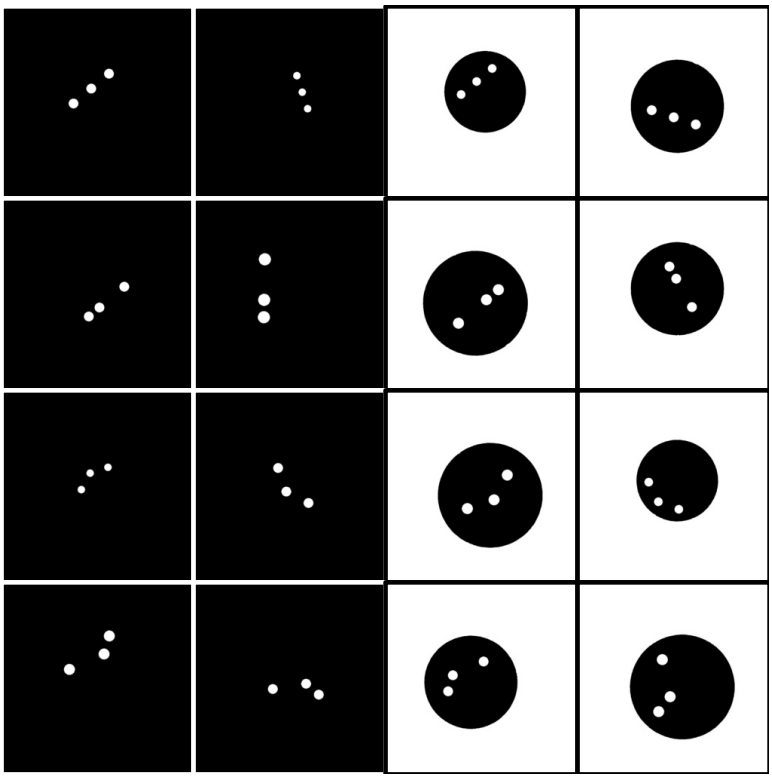

Figure 2.2: Example stimuli from each of the four test conditions. Rows depict category membership. Columns show example stimuli from each of the four test conditions. Condition 1 is identical to the training condition. Actual stimuli varied with respect to the overal scale and position of the dots within the image. For categories 2 & 4, the orientation of the implicit major axis of the dots varied from trial to trial.

differences between training angles were found, and all results are collapsed across this dimension. The beginning of each trial was signaled by the word "Ready!" onscreen. Subjects were shown the current stimulus after a key press, and were free to mouse click their response on one of four colored squares, each of which signified the "name" of each category. Assignment of color names to each category was randomized between subjects. Feedback was given for correct and incorrect decisions during training, and took on several forms. After making a choice, the correct answer was shown in color and distractors were grayed out. Correct decisions were confirmed by a smiley face displayed over the chosen answer, along with an increase in accumulated bonus points (visible onscreen at all times). Incorrect responses elicited a beep from the computer, along with a frowning face displayed over the chosen answer. Prior to training, subjects were only informed that there would be stimuli from four different categories. No information about class membership was revealed, and so participants were forced to learn to categorize our stimuli through feedback alone.

During the test period, participants were no longer given feedback of any form. All answers were confirmed by a neutral face, and bonus points were displayed as "XXXXX" onscreen. All other procedures remained the same as the training period. Within each test block, subjects were tested on an equal number of four conditions in random order (Figure 2.2). Condition 1 was identical to the training condition: white dots on a black background at an angle identical to training. For condition 2, the implicit major axis of the dot arrangements rotated freely from trial to trial, taking on any integer value between $1 - 180^\circ$, except for those values within $\pm 45^\circ$ of the participant's original training angle. In condition 3, dot patterns were embedded in a black circular context against a white background and the angle was fixed at the training direction. Condition 4 featured white dots within a black circular context and

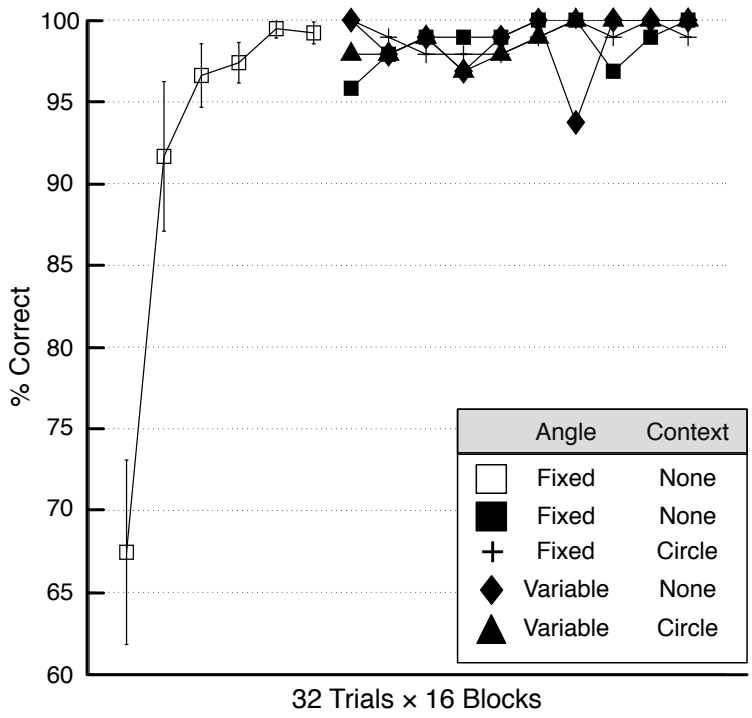


Figure 2.3: Mean accuracy across training and test for n=12 subjects. Open symbols represent training. Filled symbols show test results for conditions 1–4.

the same rotation possibilities as condition 2. When a circular context was present (conditions 3 & 4), the position of the dot arrangements varied within the circle.

### 2.1.3 Results

Mean accuracy across training and test blocks is shown in Figure 2.3 for all subjects (n=12). Performance quickly reached near-perfect asymptotic levels during training, and remained consistently high across all forms of test variability. The test phase consisted of four conditions, each designed to impose a contextual or rotational manipulation on the stimulus. Serving as a control, condition 1 was identical to the training condition. Condition 2 tested participants' resilience to stimulus rotation.

Condition 3 featured a simple change in background context, with no change in the relevant stimulus characteristics (i.e., the dots). Combining both rotation and context changes, condition 4 represented the most difficult challenge. Despite these modifications, classification accuracy remained robust across all test conditions for all of our subjects (mean accuracy of $99\% \pm 0.02$).

These results held for all of our subjects. Figure 2.4 shows mean accuracies for all twelve subjects across training and testing blocks. Despite initial naïvety, all participants demonstrated quick learning of categorical membership. In all cases, performance reached near-perfect asymptotic levels before the beginning of the test phase. For several subjects, these accuracy levels were reached by the second block of training ($\leq 8$ exposures to exemplars from each category).

Our participants demonstrated that the addition of extraneous information in the form of orientation and context variability are trivial to ignore. Presumably this is because observers are able to extract out the configural "rules" that govern class membership, and then apply these rules to new stimuli despite novel variability. By doing so, generalization across these two dimensions is obtained for free. To follow up on these results, we implemented a second study to determine if deformations in the diagnostic components (i.e., the dot arrangements) would degrade classification performance.

## 2.2 Experiment 2

### 2.2.1 Participants

Eleven individuals from The Ohio State University participated for course credit. All subjects were verified to have normal or corrected-to-normal vision by means of a Snellen eye chart.

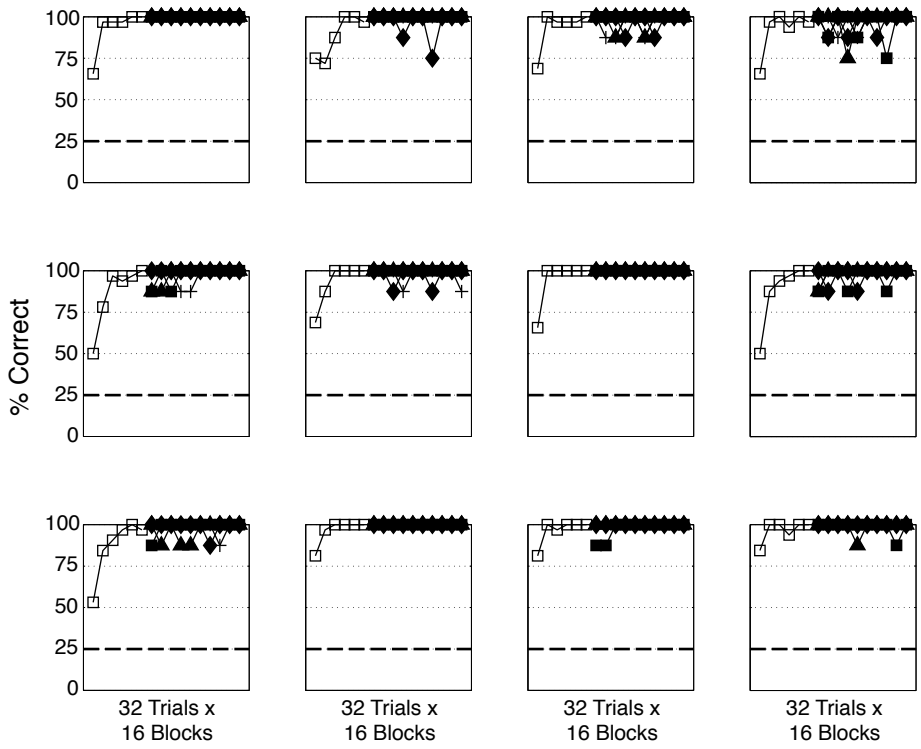


Figure 2.4: Mean accuracy on Experiment 1 across blocks for all 12 subjects. Open squares represent the training condition. Beginning at block 7, test conditions 1–4 are represented as filled squares, +, filled diamonds, and filled triangles respectively. All subjects reached near-perfect performance in relatively few blocks. Subsequent testing on four different conditions involving rotation and context generalizations had no effect on accuracy. Within blocks, each condition was tested eight times, meaning that the frequent performance levels of 87.5% are due to a single miss for that condition. Chance performance was at 25%

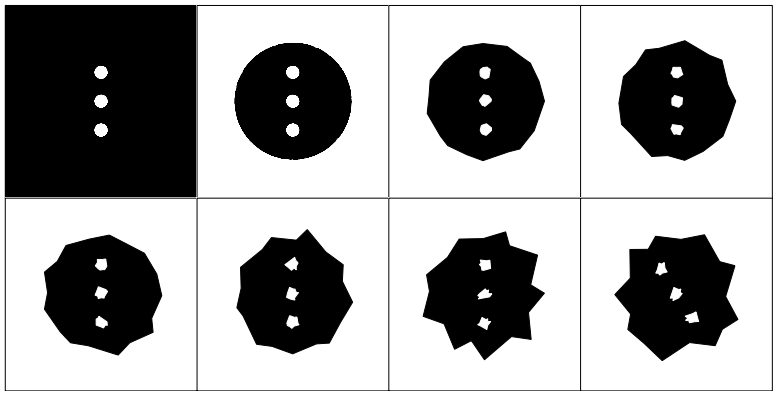

Figure 2.5: Example stimuli from Experiment 2. Stimulus variability increased across training, including the addition of a circular context, mounting irregularity in contours, and more diverse orientations.

### 2.2.2 Stimuli and Procedure

Stimuli consisted of four categories with defining properties identical to those described in Experiment 1. The task, apparatus, and experimental method also mimicked that of Experiment 1, with the following exceptions. Experiment 2 spanned two sessions on two separate days, and consisted of 20 blocks of 20 trials each per session. During session 1, subjects were trained to learn the classification task in the face of increasing stimulus variability. To begin, observers completed six training blocks composed of stimuli identical to the training stimuli in Experiment 1. This was followed by one block of stimuli featuring a black circular context. Over the next six blocks, an increasing amount of variability in the form of jagged edges was gradually introduced to the contours of dots and the circular context (see Figure 2.5). By adding irregular contours, the precise distances and positions between displaced dots were obfuscated by inconsistent deviations in the edges. This added increased complexity to the stimulus set, and made judgments regarding the absolute positions of dots difficult. Finally, variability in orientation was gradually increased across the

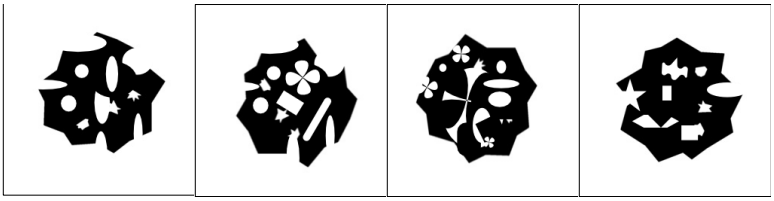

Figure 2.6: Example masks used during Experiments 2 & 3.

remaining seven blocks. In total, subjects were exposed to a subset of 150° of the total 180° of possible orientations during training. A 30° wedge of angles located in the middle of the training space was left for an interpolation test on session 2. During session 1 all stimuli were presented for 3000 msec followed by a mask. Example masks are illustrated in Figure 2.6.

Session 2 featured two components: a speed-up period, followed by a test phase. Stimuli contained the full complexity (i.e., jagged contours and orientation variability) for the entire session. Speed-up occurred during the first ten blocks, during which the stimulus presentation time was gradually reduced across blocks from 3000 msec to 250 msec. The remaining ten blocks were reserved for testing. The test procedure was similar to Experiment 1, and consisted of two conditions. In condition 1, stimuli orientations were drawn from the distribution of possible angles that subjects were exposed to during training. For condition 2, subjects were tested on orientations drawn from the 30° wedge of untrained angles.

### 2.2.3 Results

Group averaged data are shown in Figure 2.7. As in Experiment 1, our subjects were able to easily learn to identify and categorize the four classes of stimuli. Once again, the addition of a circular context, this time very early in training, had no impact

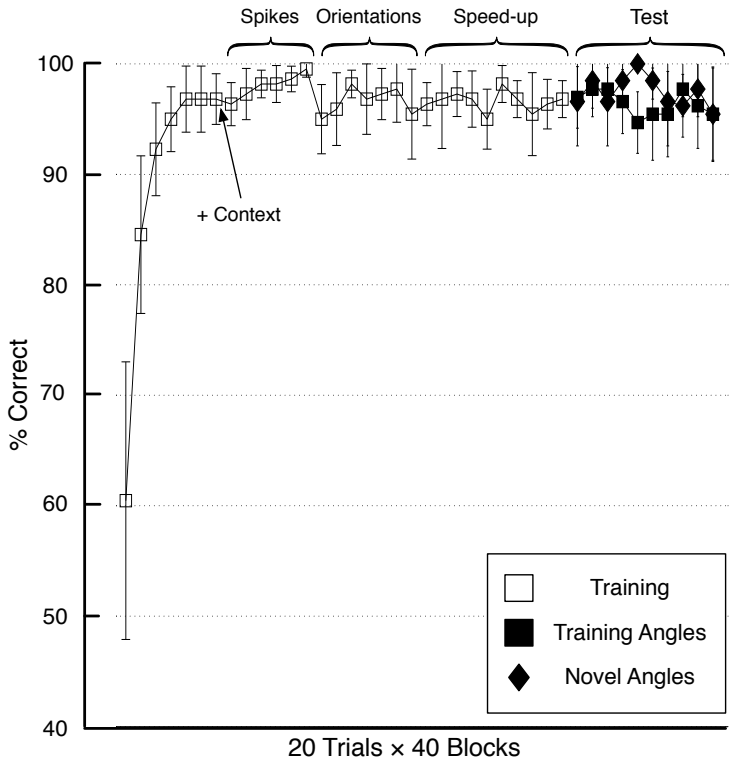


Figure 2.7: Group averaged data (n=11) from Experiment 2. Open symbols represent the training and speed-up components of the study. Filled symbols show performance during the test phase on trained orientations (squares) and novel orientations (diamonds).

on performance. Likewise, all subjects once again demonstrated robustness in the face of changes in stimulus orientation. This included performance on orientations explicitly trained on, and on a subset of interpolated angles introduced during testing. Furthermore, despite adding a large amount of irregularity to the image, the jagged contours had no effect on our participants' accuracy.

Because our subjects were able to quickly learn this task on both Experiments 1 & 2, it is likely that category boundaries were learned before the introduction of contour and orientation variability. As we have already demonstrated, observers are quite adept at ignoring extraneous details once the task has been learned. Although

the data show that people are also able to ignore contour variability, it is unclear from these results if they would be able to learn the initial categorization task with this variability present from the first trial. For our third experiment, we tested this question directly.

## 2.3 Experiment 3

### 2.3.1 Participants

Seven individuals from The Ohio State University participated for course credit. All subjects were verified to have normal or corrected-to-normal vision by means of a Snellen eye chart.

### 2.3.2 Stimuli and Procedure

The stimuli and procedure for Experiment 3 were identical to Experiment 2, with the following exceptions. First, all forms of complexity (maximally-jagged contours and 150° of possible orientations) were fully present from the first trial onward. It is important to note that the presence of orientation variability during training introduces far greater diversity in the apparent positioning of the dot arrangements, making the task far more difficult. Furthermore, the jagged contours add an element of unpredictability to the stimuli that makes the category boundaries less distinct. Because stimulus variability was present from the beginning, session 1 was used for the speed-up component of the study. Like Experiment 2, stimulus presentation time was gradually decreased from 3000 msec to a minimum of 250 msec. Stimulus presentation was followed by a mask similar to those shown in Figure 2.6. For the first half of session 2, subjects reacquainted with the task (at 250 msec presentation times).

For the remaining half, subjects were tested without feedback on the training orientations (condition 1) and on orientations drawn from the 30° wedge of interpolation angles (condition 2). Both session were composed of ten blocks of 40 trials each.

Subjects were asked to learn to classify the four stimulus categories in the face of severe deformations in the individual stimuli. That is, subjects were required to not only perceive the relationships of the dots above the noise of the jagged edges, but were forced to learn to identify what features and/or relations were diagnostic to the task. A pilot study revealed this task to be very difficult when no a priori information was given to the participants. A debriefing of our pilot subjects revealed that they were often led astray by various strategies that they employed to discover what delineates category membership. For example, several subjects reported paying too much attention to the orientation of the dot arrangements, which reveals nothing about class identity. In general, subjects reported employing a hypothesis, testing it against subsequent stimuli, and then verifying the approach via feedback. If one (or several) hypotheses failed to succeed before the correct strategy was discovered, subjects were stuck at chance performance. To help prevent such mistakes from occurring in Experiment 3, we divulged limited information to our subjects before beginning the experiment. Briefly, our participants were told that the dots were the component of the image that determined class membership. The exact verbal protocol used during instruction is given in Appendix C.

### 2.3.3 Results

The results of this experiment are summarized in Figure 2.8. Despite the hints given before training, two subjects out of seven were unable to learn the task. Debriefing revealed these subjects were once again utilizing hypothesis testing strategies for solving this problem, including such misguided strategies as counting the number of

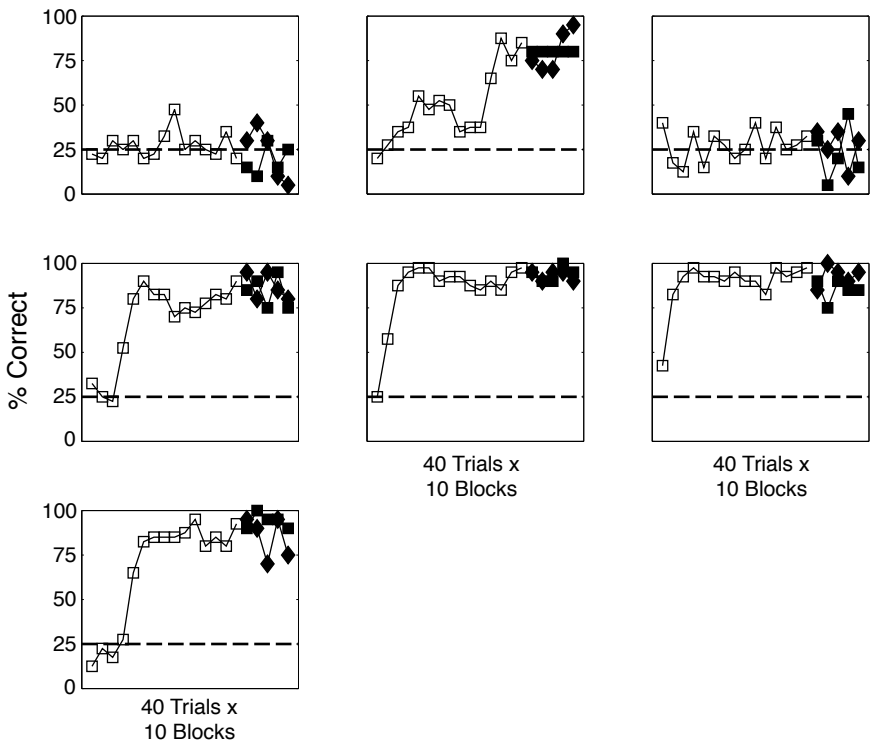


Figure 2.8: Mean accuracy for individual subject data (n=7) on Experiment 3. Open squares represent training blocks. Filled squares represent testing on the training orientations. Filled diamonds show performance on an interpolated set of orientations during testing.

spikes on the center dot in each image. This explains the poor performance of our two subjects, and the additional time-to-learning shown by our remaining five subjects relative to our previous experiments. It seems unlikely that these failures are due to an inability to encode the spatial relationships in the stimuli, but rather reflect the difficulty inherent in extracting the alignment "rules" in the face of misleading information generated by orientation and contour variability. For those subjects that were able to discover the necessary rules, an immediate "eureka" effect appears, with accuracy increasing dramatically shortly after discovery. Once again, this accuracy remained robust across testing regardless of test condition.

Taken with the results from Experiments 1 & 2, it is clear that these tasks are trivial for human observers. Additionally, our subjects demonstrated that knowledge obtained on a training corpus transfers freely across changes in orientation, background context, and contour irregularity. Furthermore, the task can be completed when stimuli are masked after fast presentation times (250 msec). Collectively, we believe that these results suggest that the configural relations of the dot arrangements are being used to complete this task. Because of this, our stimulus set could be used as a benchmark test set to determine if a particular representational structure is able to encode spatial relations. For the following computer simulations, we applied our stimulus set to a successful feature-hierarchy model of object recognition (Mutch & Lowe, 2008). If this class of models implicitly encodes spatial information as hypothesized, we would expect good performance on such an easy task.

# CHAPTER 3

# Computer Simulations

## 3.1 The HMAX Model

For model simulations, we used a version of the HMAX model (Riesenhuber & Poggio, 1999). The standard model consists of a hierarchical arrangement of five layers: an image layer of grayscale pixels, followed by four layers of alternating simple and complex cell-like units (S1, C1, S2, C2). The S1 layer is composed of predefined two-dimensional Gabor filters centered at all locations and multiple orientations. S2 prototypes are sampled randomly from C1 units during a preliminary feature-extraction stage. Template matching occurs at each S-layer, and activations are pooled through a MAX operation at each C-layer. The resulting output is a vector of C2 features that are invariant with respect to scale and position. These features are classified using an all-pairs linear support vector machine (SVM). For our work, we used the Statistical Pattern Recognition Toolbox for Matlab (Franc & Hlavac, 2004).

At the level of C2 features, where classification occurs, the standard model ignores position and scale information, reducing images to a "bag of features." To retain as much geometrical information as possible, we utilized a version of the HMAX model similar to that formulated in Mutch & Lowe, 2008 (see also, Serre, Wolf, & Poggio, 2005). The model code was made available for download at

http://www.mit.edu/ jmutch/fhlib/. Figure 3.1 illustrates the model architecture in schematic form.

Images were presented in a pyramidal fashion at ten progressively smaller scales, with a maximum size of 140x140 pixels. Grayscale images were converted by applying Gabor filters at 12 orientations to all scales and positions in each image. To reduce local clutter produced by irregular contours, we enforced sparse inputs to the S2 layer by setting within-layer inhibition levels to 50% for layers S1 and C1 (the weakest half of responses were set to zero). Together these steps suppress the activation of less dominate orientations in favor of more diagnostic ones. A limit was also imposed on the position and scale invariance of the C2 features. Conceptually, this step prevents all spatial information from being lost in the final feature vector. Algorithmically, this means that when an S2 template is being compared to a test image, the model only checks for locations in the image nearby to where the template in the original training image was originally found, where nearby means $\pm 5\%$ of the image size, and $\pm 1$ scale in the image pyramid.

## 3.2 Simulation 1

### 3.2.1 Description

Human observers are quite adept at identifying the spatial relationships associated with Vernier and bisection tasks, even under conditions requiring far more acuity than our stimuli. Because of this, and the remarkable levels of classification accuracy achieved in our experiments, it would seem that any sufficient model of vision should be able to similarly make use of such information. Our first question then was whether or not the HMAX model would be capable of achieving comparable performance on our stimuli. To determine this, we implemented a simulation that mimicked the

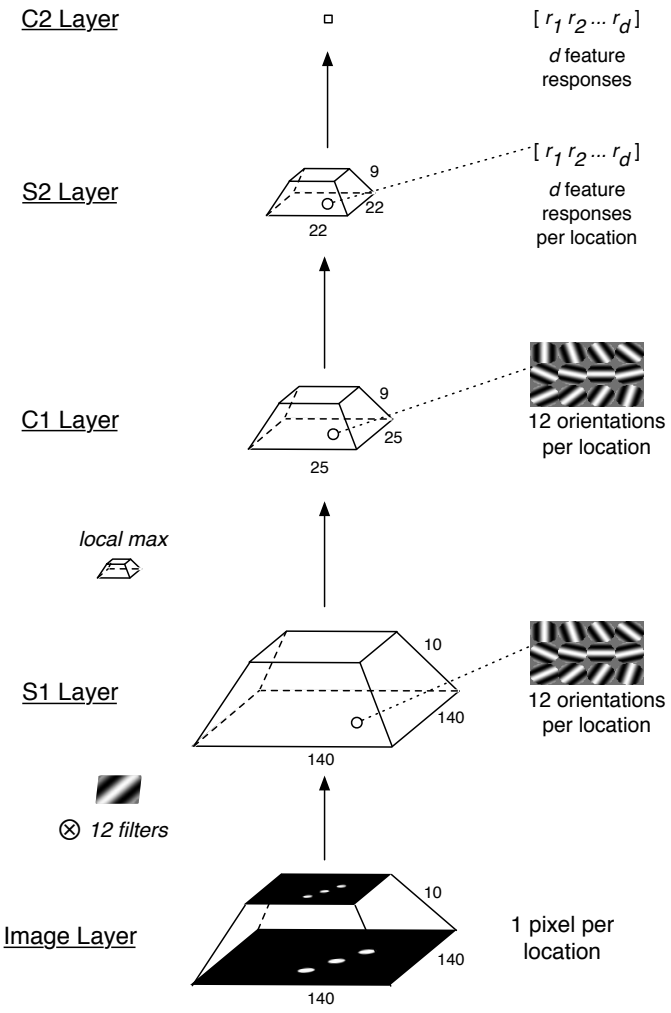


Figure 3.1: Schematic layout of the model used in our simulations (adapted from Mutch & Lowe, 2008)

design of Experiment 1. First, a model was trained on a corpus of images with properties identical to those used in the training component of Experiment 1. The training corpus included 60 images from each object class (240 images total). After training, the model was tested on each of the four test conditions. For all but test condition 1 (which was identical to the training condition) the model would be forced to generalize with respect to background context and/or orientation, just as our human participants had. All results shown are the average of 10 such runs.

### 3.2.2 Results

We tested the HMAX model of object recognition on the same four conditions used in the human experiment. Figure 3.2 compares model performance to our human observers. For condition 1 (the training condition, and left-most bar), the model achieves quite remarkable performance ($95\% \pm 0.85$). However, success on this condition requires resilience to changes in scale and position—two factors that this class of models is designed to cope with. Accuracy on conditions 2–4, shown in the middle grouping of bars in Figure 3.2 was markedly worse ($32\% \pm 6$, $43\% \pm 6$, and $26\% \pm 7$ respectively), with near-chance performance on conditions 3 and 4 when novel rotations were introduced. In fairness, the model admits to weak invariance with respect to rotation. It should be noted, though, that our human observers were able to transfer learning to new orientations with no cost, demonstrating that a significant component of representational structure is absent from this class of models.

Of particular interest are the results from test condition 2 (the shaded blue bar in the middle grouping of Figure 3.2). Stimuli for this test condition differed from the training corpus only with respect to the background context. That is, the stimulus components relevant to the task (the three dots) remained fixed at the same orientation, and varied on the dimensions of scale and position only. The only difference

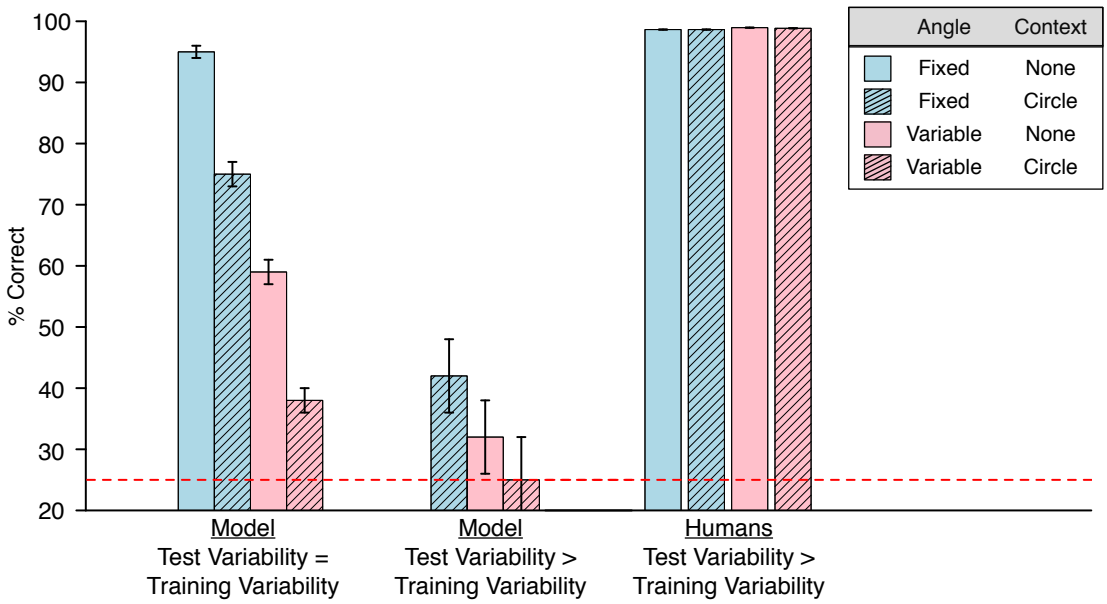


Figure 3.2: Mean accuracy of model performance for Simulations 1 & 2. The left-most grouping of bars shows model performance on Simulation 2. Individual bars, from left to right, show performance on each of the four test conditions from Experiment 1. The middle grouping of bars reflect generalization performance on conditions 2–4 of Simulation 1. Human data: mean performance across every test block for all subjects on Experiment 1. Each model data point represents the average of ten simulations ($200 \times 4 = 800$ trials per simulation). Error bars are 90% confidence intervals.

was the addition of the circular black context. We believe that such a striking drop in performance is due to the addition of novel features generated by the circular context. If the information contained in the spatial relations of the dots was being used to drive classification, we would not expect such a drop in accuracy.

## 3.3 Simulation 2

### 3.3.1 Description

For our second simulation, we wanted to know whether this model is capable of accurately classifying these stimuli under the best conditions. To address this issue, we trained four separate instantiations of the model on each of the four test conditions from Experiment 1. That is, for each test condition, a model was first trained and then tested on stimuli drawn from the same distribution of properties. This strategy helped to remove any performance deficiencies due solely to generalization weaknesses. We also tested each trained model on the other combinations of context and orientation. Note that these additional tests do require generalizations in some cases. They are included here for completeness. Stimuli were constructed identically to those used in Experiment 1. Training datasets consisted of 60 randomly generated examples from each class (240 images total). For each test condition, the model was shown $200 \times 4 = 800$ images. All results shown are the average of 10 such runs.

### 3.3.2 Results

The left-most grouping of bars in Figure 3.2 summarize the main results of Simulation 2. Again, we see that in the simplest case, the training condition from Experiment 1 (the left-most bar), the model achieves human-like performance. However, the mere addition of a circular context to the training images results in a decrease of

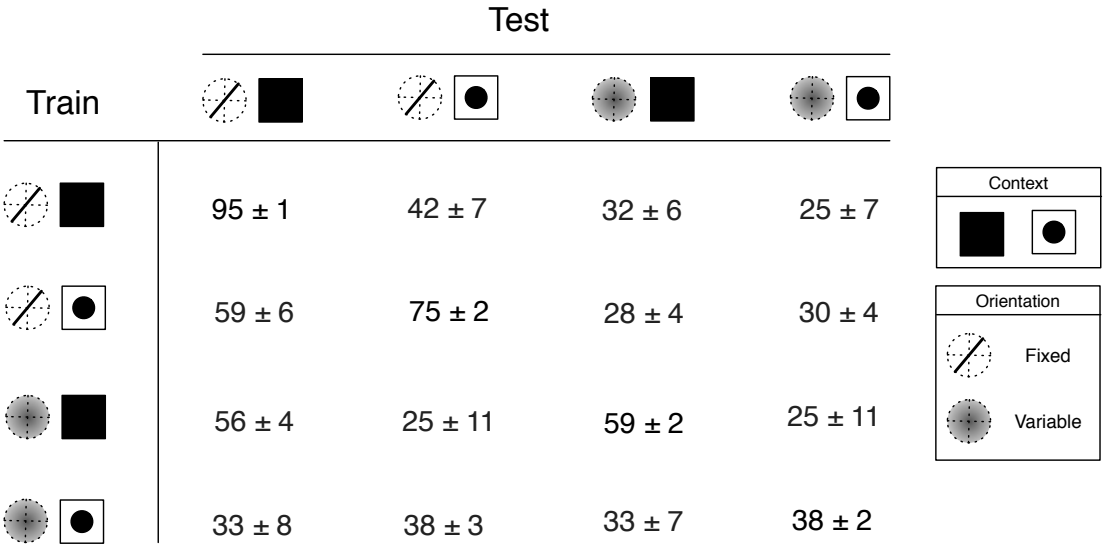


Table 3.1: Model performance from Simulation 2. Each row represents a model that was trained on a particular combination of background context and orientation(s). Columns show mean accuracy when tested on each of the possible combinations of context and orientation.

$\approx$ 20 percentage points. We see a sightly larger drop when orientation variability is added ($\approx$ 30 percentage points), and strikingly low performance with both orientation variability and circular context added to the stimuli. A possible explanation for these results will be discussed later, but it would appear that the model's limited success is due to a select few diagnostic features, not to any rule-like patterns derived from spatial configurations per se.

Table 3.1 shows the results of the entire simulation for the test conditions (along the diagonal) and all other possible generalization tests. In general, two trends can be identified from the data. First, as stimulus complexity increases in the training corpus, model classification performance decreases. Second, decreased stimulus complexity in the training corpus leads to poorer generalization performance. The

latter effect is likely due in part to overfitting. Unfortunately, avoiding this problem requires increasing the number of training items to at least the number of items in the feature vectors used for classification (4075), which is currently an intractable problem. Nonetheless, even under a sufficient training corpus we should not expect generalization performance to surpass accuracy levels on the test corpus. This restricts theoretical performance to at least the levels shown along the diagonal of Table 3.1.

## 3.4 Simulation 3

### 3.4.1 Description

Simulations 1 and 2 revealed that our classification task becomes increasingly more difficult for the model as stimulus complexity increases. Subjects in Experiment 2 made it clear that adding such complexity, in the form of jagged contours, rotation, and faster presentation times, does little to degrade classification in human performers.

There is no obvious analogous method for training the model on incremental complexity as in Experiment 2. To address this, we devised a method of approximating such a training paradigm. We began by first creating eight steps, each defined by the complexity of stimuli contained within. Step 1 was composed of stimuli featuring circular dots against a uniform black background. A circular context was introduced at step 2. Steps 3–5 introduced increasing jaggedness to dot and circular context contours. Finally, steps 6–8 gradually introduced orientation variability, up to a full range of 180° at step 8.

Next, we trained eight individual models on each of the eight steps, such that at each step, the training corpus featured a mixture of stimuli from that step and all

| Step # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Step 1 | 60 | | | | | | | |
| Step 2 | 30 | 30 | | | | | | |
| Step 3 | 10 | 20 | 30 | | | | | |
| Step 4 | 5 | 10 | 15 | 30 | | | | |
| Step 5 | 5 | 5 | 10 | 10 | 30 | | | |
| Step 6 | 5 | 5 | 5 | 5 | 10 | 30 | | |
| Step 7 | 5 | 5 | 5 | 5 | 5 | 5 | 30 | |
| Step 8 | 5 | 5 | 5 | 5 | 5 | 5 | 10 | 20 |

Table 3.2: Training corpus breakdown for the eight steps in Simulation 3. For each step, the training image set contained a mixture of images from the current step and all previous steps. For example, the Step 3 training set contained 10 stimuli that fit the Step 1 description, 20 from Step 2, and 30 from step 3.

previous steps (for a total of 60 images per class). For example, the model trained at step 2 had a training corpus of 30 stimuli with no background context and 30 stimuli with a circular context (from each class). Table 3.2 shows the exact training stimuli mixture for each step.

After training, each model was then tested on stimuli from each of the eight steps. Importantly, test corpora contained only images from each individual step. That is, for step 2 test images, all images contained stimuli with a circular background context, smooth contours, and a fixed orientation. For each test step there were 200 images per class, and all reported results are the average of ten repetitions.

### 3.4.2 Results

The results of the step simulation are summarized in Figure 3.3. For models trained

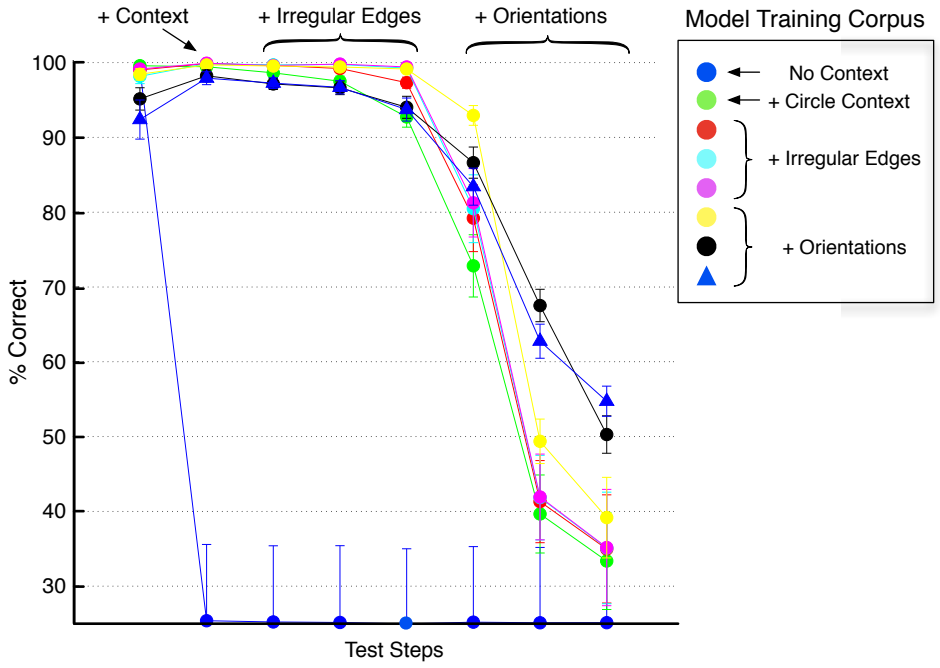


Figure 3.3: Model accuracy on Simulation 3. Each line represents an individual model that was trained on a unique mixture of stimuli. Training corpora become increasingly complex at each step (see Table 3.2). Each model was tested on all of the possible test "steps" (shown along the x-axis).

on any step except for step 1 (in which stimuli had uniform black backgrounds), increases in contour spikiness (points 1–5 on the x-axis) present little trouble. This success could be a result of the sparsification method we employed in the model, or due to blurring occurring in the early layers of the model, which would smooth over irregularities in dot contours (see discussion). With the introduction of rotation, we see a significant drop in performance (points 6–8). There is, however, a substantial improvement in training the model in this fashion over training solely on stimuli with full complexity only. For instance, the model trained on step 8 (blue triangles) saw comparable images to the model in the final row of Table 3.1. However, the step 8 training corpus also had images featuring less variability (drawn from steps 1–7). This mixture raises the performance of the model from 38% to 55%. Importantly, these performance increases are not due entirely to the training mixture. For both the training and test stimuli in this simulation, the position of the stimulus was restricted to the lower-left quadrant of the image. This reduced the possible problem space, simplifying the task considerably. Regardless, even with the beneficial mixture of training images, peak performance remained remarkably low, reflecting once again a deficiency in the coding scheme used by the model.

# CHAPTER 4

# Discussion

All things considered, we see two very different strategies emerging from our human and model results. The pattern of learning we see in the individual subject data reveals a point of "eureka-like" learning, wherein performance increases radically over a very short time period. Unlike incremental learning, this trend suggests the discovery and adoption of category rules that, once obtained, can be applied with consistent accuracy across a range of conditions. Furthermore, the ability of our observers to continue to successfully categorize our stimuli under various forms of transformations strengthens the hypothesis that they are using the arrangement rules to complete the task, rather than one or more individual features in isolation. In fact, when asked, our subjects invariably described the four categories in terms of the relative spatial positions of the three dots.

By contrast, the model only manages to achieve acceptable performance under a limited range of conditions. Whatever representation the model is using is clearly different than what human observers use. This raises the interesting question of what information the model is taking advantage of, and how it is able to achieve such good performance on condition 1 in Simulation 1. Remember that under this condition, the dot arrangements always fell along a fixed orientation. By varying the position and scale of the stimuli from trial to trial, we prevented the model from solving the task by simply noting the presence or absence of dots in specific locations. One likely

answer is that the model was using a lowpass-filtered "feature" that encompassed the entire dot arrangement as a single entity. By doing so, the blurred dot patterns could be reduced to straight, or one of several curvilinear segments. In fact, to achieve the level of performance listed in our results, we first had to perform a parameter search to find optimized settings for the model (see Appendix B). The one value that produced the best performance boosts was the receptive field (RF) size of the S1 layer Gabors. This search led us to increase the RF size of each filter from 11x11 pixels to 27x27 pixels. At the smaller scales in the image pyramid, this results in Gabor filters that encompass the entire image. In a previous version of the HMAX model (Serre et al., 2005), images were convolved with a battery of Gabors set at varying receptive field sizes ranging from 7x7 (0.19° visual angle) to 39x39 (1.07° visual angle) in steps of 2 pixels. These values were chosen in Serre & Riesenhuber, 2004 to be consistent with properties of parafoveal simple cells (Schiller, Finlay, & Volman, 1976). The version of our model by Mutch & Lowe, 2008 uses a different method for approximating these receptive fields. Rather than using multiple filter sizes, the same size filter is applied to the images at different scales. By increasing the RF size in the way we have (to give the model the best chance possible), we have likely added a biologically unrealistic component of blurring which has helped the model significantly.

In another experiment not listed above we find more evidence that supports our hypothesis. To push the limits of our observers, we trained several subjects to classify our stimuli under extremely difficult conditions. From the first trial, stimuli contained maximum contour irregularities, full orientation variability, and stimuli were masked after 250 msec presentation times. In addition to this, the Vernier and bisection displacements were reduced, making their relative positions very difficult to identify when combined with the jagged contours. Subjects were given no a priori information

beyond the number of categories. The experiment and data are not presented in full because the task proved to be too difficult for half of our subjects. For those subjects that failed, all once again reported failing due to the hypothesis testing strategy (i.e., time was wasted focusing on irrelevant features such as contour spikes, orientations, etc.). Several highly motivated graduate students were able to succeed at the task. We found that their accuracy remained at chance until the rule was learned, whereupon their performance immediately increased to near-ceiling levels. Despite initial difficulties, it appeared again that our subjects were using rule-like templates based on the configural relations of individual features.

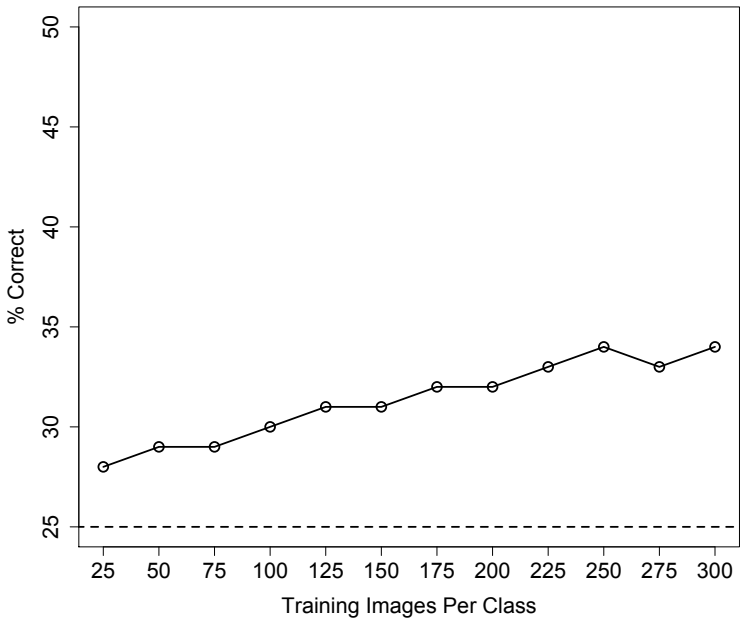


Figure 4.1: Model performance on stimuli with reduced Vernier and bisection displacements. With smaller dot displacements and jagged contours, blurred stimuli from the four classes become more homogenous in appearance. Tremendously poor performance, even when the number of training images per class is increased to 300, suggests that the model relies on blurred "global" features to achieve limited success.

We then trained and tested the model on the same stimuli. If the model is using blurring to treat the entire dot arrangements as single features, then we would expect the coupling of finer Vernier/bisection displacements with the jagged contours to produce less diagnostic global features, and thus worse performance. This seems to be the case, as Figure 4.1 illustrates. Despite increasing the number of training images from 25 up to 300 per class, performance remained near chance. By reducing the displacement of the central dot, the blurred dot arrangements begin to appear more collinear, regardless of class. The decrease in model accuracy suggests that this is in fact what the model is using to achieve its limited success.

Taken together, these results once again suggest that this class of models use a particular, limited strategy for solving recognition problems. That is, a collection of free-floating features are being used to identify each image's content. It seems unlikely that these features have implicitly encoded the relevant spatial configurations of object parts via overlapping redundancy. In most applications, such as multi-class recognition tasks on the Caltech 101 image database (Fei-Fei, Fergus, & Perona, 2004), this approach is sufficient for success. With the more complex objects contained within these databases, a mere mixture of unlocalized features can reliably differentiate tigers from motorcycles or elephants from zebras. However, while this strategy works in some scenarios, it is not necessarily reflective of what information human observers use to complete the same tasks. We have created a simple stimulus set based on spatial relations fundamental to human vision. Our results suggest that while humans can readily discriminate between these categories, feature-hierarchy approaches to object recognition lack a critical component necessary for utilizing spatial configurations of objects and object parts.

We cannot stop at the failure of these models alone. The specific deficiencies found in this class of models should push us toward efforts for improvement. One

natural step is to look to modeling research on Vernier stimuli. Poggio et al., 1992 demonstrated that a simple hyper basis function (hyperBF) network could quickly learn to perform the Vernier task at hyperacuity levels using units with overlapping receptive fields. In this approach, a hidden layer containing orientation-tuned V1-like units with elongated receptive fields processes a layer of overlapping inputs. Given training, a decision unit downstream integrates over the hidden layer to determine the offset direction of the Vernier stimulus. However, the model used hand-coded photoreceptor cell-like units and was meant to be little more than a proof of concept. Later variations (Weiss, Fahle, & Edelman, 1993; Sotiropoulos, Seitz, & Seriès, 2011) would extend this approach by adding an initial layer of oriented Gabor filters capable of processing raw images, along with more sophisticated learning mechanisms. As the focus of these models is typically on learning to improve on hyperacuity tasks, these models lack the broader applicability of more general-purpose feature hierarchy models. Furthermore, the low-level V1-like assumptions driving these models are not unlike those used in early levels of the HMAX model. This means that they will capture simple, oriented features with little-to-no regard for spatial position, once again creating a data structure unsuitable for differentiating stimuli such as ours.

Along with the hyperBF class of models, an emerging list of models of spatial vision have emerged capable of completing simple Vernier and bisection acuity tasks (e.g., Geisler & Super, 2000; Zhaoping, 2003; Thielscher & Neumann, 2003). Once again, we find a reliance on a similar foundation of low-level, oriented feature detectors devoid of spatial information. Another approach, known as Wilson-Cowan Type Models (WCTM), has emerged in parallel, and has also demonstrated proficiency on Vernier stimuli. These models (Hermens, Luksys, Gerstner, & Herzog, 2008; Rüter, Francis, Frehe, & Herzog, 2011), built upon the work of Wilson & Cowan, 1973,

function as irregularity detectors. WCTM approaches typically involve a simple two-layer network wherein redundant features are suppressed through lateral inhibition, while irregularities are enhanced. Although this class of models has been successful at Vernier tasks under a variety of conditions, they are merely simple mathematical demonstrations targeting a very limited scope of phenomena. It is not clear if or how this approach could scale up to include orientation, size, and position variability.

Other approaches are also being used to model Vernier phenomena. Some very promising models are coming out of research on figure-ground segmentation, and perceptual grouping research. The 3D LAMINART model (Cao & Grossberg, 2005; Francis, 2009) is a powerful and general model of vision and has been applied to many diverse visual tasks including, but not limited to Vernier hyperacuity. Other similar approaches, such as the work of Craft, Schütze, Niebur, & von der Heydt, 2007, present other exciting possibilities relying upon similar modeling assumptions. The details of these models are beyond the scope of this work, and it is currently unclear whether such modeling strategies will offer sufficient solutions to the current problem posed in this work, but they will certainly bring new insights that may shed additional light on what is necessary for encoding relational information.

## References


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

Bar, M., Tootell, R. B., Schacter, D. L., Greve, D. N., Fischl, B., Mendola, J. D., et al. (2001). Cortical mechanisms specific to explicit visual object recognition. *Neuron*, *29*(2), 529–535.

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy & R. Pomerantz (Eds.), *Perceptual organization* (pp. 213–263). Hillsdale, NJ: Erlbaum.

Biederman, I. (1987). Recognition-by-components; a theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Cao, Y., & Grossberg, S. (2005). A laminar cortical model of stereopsis and 3D surface perception: Closure and da vinci stereopsis. *Spatial Vision*, *18*, 515–578.

Craft, E., Schütze, H., Niebur, E., & von der Heydt, R. (2007). A neural model of figure-ground organization. *Journal of Neurophysiology*, *97*, 4310–4326.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition*, 886–893.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341.

Fahle, M., & Morgan, M. (1996). No transfer of perceptual learning between similar

stimuli in the same retinal position. *Current Biology*, *6*(3), 292–297.

Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *CVPR workshop on generative-model based vision*.

Franc, V., & Hlavac, V. (2004). *Statistical pattern recognition toolbox for Matlab.*

Francis, G. (2009). Cortical dynamics of figure-ground segmentation: Shine-through. *Vision Research*, *49*, 140–163.

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

Geisler, W. S., & Super, B. J. (2000). Perceptual organization of two-dimensional patterns. *Psychological Review*, *107*(4), 677–708.

Green, C., & Hummel, J. E. (2004). Functional interactions affect object detection in non-scene displays. In K. Forbus, D. Gentner, & T. Reiger (Eds.), *Proceedings of the 26th annual conference of the cognitive science society* (pp. 488–493). Mahwah, NJ: Erlbaum.

Green, C., & Hummel, J. E. (2006). Familiar interacting object pairs are perceptually grouped. *Journal of Experimental Psychology*, *32*, 1107–1119.

Hayworth, K. J., Lescroart, M. D., & Biederman, I. (2010). Visual relation encoding in anterior LOC. *Journal of Experimental Psychology: Human Perception and Performance*, *(Epub ahead of print)*.

Hermens, F., Luksys, G., Gerstner, W., & Herzog, M. H. (2008). Modeling spatial and temporal aspects of visual backward masking. *Psychological Review*, *115*(1), 83–100.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape perception. *Psychological Review*, *99*(3), 480–517.

Humphreys, G. W. (1987). *Visual object processing: A cognitive neuropsychological approach*. Hove, United Kingdom: Lawrence Erlbaum Associates.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*(2), 175–219.

Kim, J., & Biederman, I. (2010). Where do objects become scenes? *Cerebral Cortex*, *(Epub ahead of print)*.

Kim, J. G., Biederman, I., Lescroart, M. D., & Hayworth, K. J. (2009). Apaptation to objects in the lateral occipital complex (LOC): Shape or semantics? *Vision Research*, *49*, 2297–2305.

Klein, S. A., & Levi, D. M. (1985). Hyperacuity thresholds of 1 sec: Theoretical predictions and empirical validation. *Journal of the Optical Society of America, A2*, 1170–1190.

Leung, T. K., & Malik, J. (2001). Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Conference on Computer Vision*, *43*(1), 29–44.

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology*, *20*(5), 1015–1036.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. *International Conference on Computer Vision*, 1150–1157.

The MathWorks. (2009). MATLAB user's guide [Computer software manual]. Natick, MA: The MathWorks, Inc.

Mutch, J., & Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision*, *80*(1), 45–57.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive*

*neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Perret, D. I., & Oram, M. (1993). Neurophysiology of shape processing. *Image and Vision Computing*, *11*, 317–333.

Poggio, T., Fahle, M., & Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science*, *256*, 1018–1021.

Pylyshyn, Z. W. (2007). *Things and places: How the mind connects with the world*. Cambridge, MA: MIT Press.

Pylyshyn, Z. W. (2009). Perception, representation, and the world: The FINST that binds. In D. Dedrick & L. Trick (Eds.), *Computation, cognition, and Pylyshyn* (pp. 3–48). Cambridge, MA: Bradford.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Rüter, J., Francis, G., Frehe, P., & Herzog, M. H. (2011). Testing dynamical models of vision. *Vision Research*, *51*, 343–351.

Schiele, B., & Crowley, J. L. (1996). Probabilistic object recognition using multidimensional receptive field histograms. *International Conference on Pattern Recognition*, *B*, 50–54.

Schiller, P. H., Finlay, B. L., & Volman, S. F. (1976). Quantative studies of single-cell properties in monkey striate cortex i. spatiotemporal organization of receptive fields. *Journal of Neurophysiology*, *39*(6), 1288–1319.

Seitz, A. R., Yamagishi, N., Werner, B., Goda, N., Kawato, M., & Watanabe, T. (2005). Task-specific disruption of perceptual learning. *Proceedings of the National Academy of Sciences*, *102*(41), 14895–14900.

Serre, T., & Riesenhuber, M. (2004, November). *Realistic modeling of simple and complex cell tuning in the HMAX model, and implications for invariant object recognition in cortex* (CBCL Paper 239 / AI Memo 2004-107). Cambridge, MA: Massachusetts Institute of Technology.

Serre, T., Wolf, L., & Poggio, T. (2005). Object recognition with features inspired by visual cortex. *Computer Vision and Pattern Recognition*, 1–7.

Sotiropoulos, G., Seitz, A. R., & Seriès, P. (2011). Perceptual learning in visual hyperacuity: A reweighting model. *Vision Research*, *51*(6), 585–599.

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, *7*(1), 11–32.

Thielscher, A., & Neumann, H. (2003). Neural mechanisms of cortico-cortical interaction in texture boundary detection: A modeling approach. *Neuroscience*, *122*, 921–939.

von der Malsburg, C. (1999). The what and why of binding: The modeler's perspective. *Neuron*, *24*(1), 111–125.

Waugh, S. J., & Levi, D. M. (1995). Spatial alignment across gaps: Contributions of orientation and spatial scale. *Journal of Optical Society of America*, *12*, 2305–2317.

Weiss, Y., Fahle, M., & Edelman, S. (1993). Models of perceptual learning in vernier hyperacuity. *Neural Computation*, *5*, 695–718.

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

Westheimer, G. (1981). Visual hyperacuity. *Progress in Sensory Physiology*, *1*, 1–37.

Westheimer, G., Crist, R. E., Gorski, L., & Gilbert, C. D. (2001). Configuration specificity in bisection acuity. *Vision Research*, *41*(9), 1133–1138.

Wilson, H. R., & Cowan, J. D. (1973). A mathematical theory of the functional

dynamics of cortical and thalamic nervous tissue. *Kybernetik*, *13*, 55–80.

Zhaoping, L. (2003). V1 mechanisms and some figure-ground and border effects. *Journal of Physiology, Paris*, *97*, 503–515.

# Appendix A: Unix Cluster Environment

All of the above model simulations were carried out on the IBM Unix cluster "Glenn" at the Ohio Supercomputer Center. The cluster consists of 1629 compute nodes consisting of 2.5 and 2.6 GHz AMD Opteron quad core processors running with 8 to 64 GB of RAM per node. The entire cluster is fully connected by a 10 Gbps Infiniband ConnectX host channel adapter (HCA), providing performance levels greater than 22 trillion floating point operations per second.

Model simulations described below were run using a combination of Matlab and optimized C scripts. Batch jobs were carried out on the cluster by use of the TORQUE resource manager and the Moab Scheduler. Despite using sequential, single-CPU scripting, the cluster allowed for execution of simulations in parallel, with work spread out over multiple nodes, utilizing the combined strength of multiple processors and large amounts of RAM. This powerful setup allowed for computationally heavy operations, such as parameter searches, to be reduced from months of processing time down to several days. Batch processing significantly reduced even the simplest of simulations. For example, for a relatively small job, such as training our object recognition model on relatively few images (50–60 images per category), typical processing times lasted approximately 2–3 hours. To obtain stable estimates of performance, all of our simulations were replicated 10 times per test condition. This amounts to 20–30 hours of sequential processing time *per data point*. For most simulations, there were 4+ test conditions, bringing total processing time to 80–120 hours even in this simple case.

With parallel batch processing, total simulation times are restricted to the minimum time associated with a given parameterization of the model. In this example, the entire 80–120 hours of processing time can be completed in just 2–3 hours. For computationally heavy operations, such as parameter searches, that require far more, and far longer simulations to be carried out, the benefits of clustered processing become more apparent.

In addition to the hardware described above, the "Glenn" cluster provides researchers access to 18 Quadro Plex S4's, each consisting of 4 Quadro FX 5800 GPU's with 4 GB of memory per card. This arrangement provides 72 CUDA-enabled graphics devices fully connected by a 20 Gbps Infiniband ConnectX HCA, allowing for over 75 trillion floating point operations per second. Although we did not utilize this hardware for the simulations described above, our future efforts will likely involve these options, as GPU-based processing enables significant speed-ups over traditional CPU-driven processing.

## Appendix B: Parameter Search

Because our stimuli are atypical relative to those images used to optimize previous instantiations of the HMAX model, we also elected to run an optimized parameterization search.

We began by running a simulation wherein we doubled the number of features from 4075 to 8150. This slowed the simulation down tremendously and provided no substantive improvements in accuracy. The original number of features (4075) was used for the remainder of the simulations (a number commonly used in the literature (e.g., Serre et al., 2005; Mutch & Lowe, 2008)).

Next, we ran a large number of simulations, adjusting (in order): (1) receptive field size (17 values spanning from 7x7 pixels to 39x39 pixels, in steps of 2), (2) within-layer inhibition (9 equally spaced values from 0 to 1), (3) xy tolerance (7 values from 2.5% to 100%), and (4) scale tolerance (from 0 to 8 scales). Inhibition refers to what percentage of S1 and C1 unit responses are set to zero (described as sparsification above). Scale and xy tolerance levels determined what amount of position and scale invariance the model allowed in the final C2 features. Smaller numbers attempt to retain more geometrical information, while larger values boost the model's level of invariance at the cost of sacrificing spatial information contained in the image. Our method was to use cumulative tuning as in (Mutch & Lowe, 2008). That is, we first found the optimum receptive field size. We then fixed that value and it became our starting point for a search of inhibition, and so on. We averaged performance across

10 repetitions for each value of every parameter, for a total of 420 simulations. Use of the Unix cluster made such an endeavor possible.

The only parameter that appreciably affected performance was the receptive field size of the simple cell layer (S1) gabors. Setting the receptive field size to 27x27 pixels increased performance by $\approx 60\%$ over the value originally used (11x11 pixels). The remaining optimized parameters were found to be equivalent to Mutch & Lowe, 2008, so we elected to leave them as is. Table B.1 lists the final parameter values that were chosen.

| Parameter | Mutch & Lowe 2008 | Our Model |
|---|---|---|
| # Features | 4075 | 4075 |
| RF Size | 11 | **27** |
| Inhibition | 0.5 | 0.5 |
| xy Tolerance | 0.05 | 0.05 |
| Scale Tolerance | 1 | 1 |

Table B.1: Parameter values used by Mutch & Lowe, 2008, along with those adopted in our simulations as a result of our parameter search.

# Appendix C: Experimental Instructions

## Experiment 1 Instructions

*The following is to be read aloud to each participant after giving informed consent:*

- You must have normal, or corrected-to-normal vision (20/20). (Utilize the Snellen chart and note the participants level of visual acuity.)
- There are no known risks in this experiment.
- This experiment will take 1 session to complete.
- You will be shown simple images on a computer screen and be asked to classify them into four categories by clicking on the screen with the left mouse button.
- There are four categories: "Red," "Blue," "Green," and "Yellow." One image is presented on each trial and you classify it by left-clicking the appropriate color square. The colors reveal nothing about the categories themselves. We are treating the colors only as names, much in the way you call a table a "table," or a chair a "chair."
- The computer follows certain predefined procedures to generate the images from each category. Your job is to figure out what kinds of images belong to which category. The feedback given on each trial will assist you in this.

- Each trial begins with the appearance of the word "Ready!". When this appears onscreen, and you are ready to begin, press the spacebar to have this trial's stimulus presented onscreen. The stimulus then appears and stays on the screen until you enter your response. Once the stimulus has appeared, a mouse cursor will become active near the response color grid. You are free to click on the color of your choice once you are ready. The computer will continue to wait until you have done so.

- After responding, several things will occur: 1) The incorrect color squares will be grayed out, leaving only the correct answer in color, and 2) the original stimulus will reappear onscreen for a set period of time so that you can look at it once again. Additionally, if your response is correct, the bonus points will increase by one, and a smiley face will appear over the correct answer that you have chosen.

- If your answer is incorrect: You will hear a "bad" beep, and a frowning face will appear over your chosen square.

- Use this feedback as a guide in figuring out how to separate the four categories. The computer will not be playing tricks on you. It follows a consistent pattern that does not change throughout the experiment. It is possible to achieve very high accuracy.

- Accuracy is more important than speed, but your response times are recorded also. Try to respond as quickly as possible without making too many errors.

- Each session is organized in 16 blocks of 32 trials each. An equal number of "Red," "Blue," "Green," and "Yellow" images occur in every block, in random

order. The image on any particular trial is generated independently of that on any other trial.

- Todays session will be broken down into two separate periods: Blocks 1-6: Trials will proceed as described above. Blocks 7-16: You will no longer be given any feedback (neither positive or negative feedback). Your bonus points will appear as X's (XXXX) onscreen, and every response, whether correct or incorrect will be confirmed by an "indifferent" face. You will still receive bonus points, so continue to try diligently despite the lack of feedback. Some aspects of the images and the background color of the computer monitor may change from time to time, but the categories will remain the same as before.

- Two things I should let you know about the experimental room: we're going to have you place your chin on a chin rest during the experiment. This is just to make sure that your position relative to the screen remains the same. Also, the experimental room isn't very large, just to give you a heads up in case you might feel uncomfortable in a small room.

- Again, there are no known risks associated with this experiment, and all data is stripped of any identifying characteristics.

- Do you have any questions about participation?

## Experiment 2 Instructions

*The following is to be read aloud to each participant after giving informed consent:*

- You must have normal, or corrected-to-normal vision (20/20). (Please utilize the Snellen chart and note the participants level of visual acuity.)

- There are no known risks in this experiment.

- This experiment will take up to 2 separate days to complete.
- You will be shown simple images on a computer screen and be asked to classify them into four categories by clicking on the screen with the left mouse button.
- There are four categories: "Red," "Blue," "Green," and "Yellow." One image is presented on each trial and you classify it by left-clicking the appropriate color square. The colors reveal nothing about the categories themselves. We are treating the colors only as names, much in the way you call a table a "table," or a chair a "chair."
- The computer follows certain predefined procedures to generate the images from each category. Your job is to figure out what kinds of images belong to which category. The feedback given on each trial will assist you in this.
- Each trial begins with the appearance of a fixation dot. When the dot has appeared onscreen, and you are ready to begin, press the spacebar to have this trial's stimulus presented onscreen. The stimulus then appears near the point of the original fixation dot and stays on the screen for a period of time. Once the stimulus has appeared, a mouse cursor will become active near the response color grid. You are free to click on the color of your choice once you are ready. If you have not responded after the set period of time, the image will be replaced by a "mask" image. The purpose of this mask is to prevent you from seeing the stimulus any longer. You should otherwise ignore the content of the mask. You do not have to wait for the mask to appear to input your response. If you have still not responded, the computer will continue to wait until you have done so.
- After responding, several things will occur: 1) The incorrect color squares will be grayed out, leaving only the correct answer in color, and 2) the original

stimulus will reappear onscreen for a set period of time so that you can look at it once again. Additionally, if your response is correct, the bonus points will increase by one, and a smiley face will appear over the correct answer that you have chosen.

- If your answer is incorrect: You will hear a "bad" beep, and a frowning face will appear over your chosen square.

- Use this feedback as a guide in figuring out how to separate the four categories. The computer will not be playing tricks on you. It follows a consistent pattern that does not change throughout the experiment. It is possible to achieve very high accuracy.

- Accuracy is more important than speed, but your response times are recorded also. Try to respond as quickly as possible without making too many errors.

- Each session is organized in 20 blocks of 20 trials each. An equal number of "Red," "Blue," "Green," and "Yellow" images occur in every block, in random order. The image on any particular trial is generated independently of that on any other trial.

- Session 2 specific details include (wait until session 2 to read these):

- At the beginning of Session 2, both the duration of the stimulus presentation and feedback presentation will be several seconds, just as in session 1. Over the course of the second session, both presentation times will gradually reduce to shorter intervals. The task remains the same regardless.

- For the second half of session 2, you will no longer be given any feedback (neither positive or negative feedback). Your bonus points will appear as X's (XXXX) onscreen, and every response, whether correct or incorrect will be confirmed

by an "indifferent" face. You will still receive bonus points, so continue to try diligently despite the lack of feedback.

- Two things I should let you know about the experimental room: we're going to have you place your chin on a chin rest during the experiment. This is just to make sure that your position relative to the screen remains the same. Also, the experimental room isnt very large, just to give you a heads up in case you might feel uncomfortable in a small room.

- Again, there are no known risks associated with this experiment, and all data is stripped of any identifying characteristics.

- Do you have any questions about participation?

## Experiment 3 Instructions

*The following is to be read aloud to each participant after giving informed consent:*

- You must have normal, or corrected-to-normal vision (20/20). (Please utilize the Snellen chart and note the participants level of visual acuity.)

- There are no known risks in this experiment.

- This experiment will take up to 2 separate days to complete.

- You will be shown simple images on a computer screen and be asked to classify them into four categories by clicking on the screen with the left mouse button.

- There are four categories: "Red," "Blue," "Green," and "Yellow." One image is presented on each trial and you classify it by left-clicking the appropriate color square.

- The computer follows certain predefined procedures to generate the images from each category. Your job is to figure out what kinds of images belong to which category. The feedback given on each trial will assist you in this.

- Each trial begins with the appearance of a fixation dot. When the dot has appeared onscreen, and you are ready to begin, press the spacebar to have this trials stimulus presented onscreen. The stimulus then appears near the point of the original fixation dot and stays on the screen for a period of time. Once the stimulus has appeared, a mouse cursor will become active near the response color grid. You are free to click on the color of your choice once you are ready. If you have not responded after the set period of time, the image will be replaced by a "mask" image. The purpose of this mask is to prevent you from seeing the stimulus any longer. You should otherwise ignore the content of the mask. You do not have to wait for the mask to appear to input your response. If you have still not responded, the computer will continue to wait until you have done so. After responding, the incorrect color squares will be grayed out and the original stimulus will reappear onscreen for a set period of time. Additionally, if your response is correct, the bonus points will increase by one, and a smiley face will appear over the correct answer that you have chosen. If your answer is incorrect, you will hear a "bad" beep, and a frowning face will appear over your chosen square. Use this feedback as a guide in figuring out how to separate the four categories. The computer will not be playing tricks on you. It follows a consistent pattern that does not change throughout the experiment. It is possible to achieve very high accuracy.

- Accuracy is more important than speed, but your response times are recorded also. Try to respond as quickly as possible without making too many errors.

- Each session is organized in 10 blocks of 40 trials each. An equal number of "Red," "Blue," "Green," and "Yellow" images occur in every block, in random order. The image on any particular trial is generated independently of that on any other trial.

- At the beginning of Day 1, both the duration of the stimulus presentation and feedback presentation will be several seconds. Over the course of the first session, both presentation times will gradually reduce to shorter intervals. The task remains the same regardless.

- On Day 2, presentations times will be fixed for all trials at the most rapid duration from session 1. For the first half of session 2, each trial will progress identically to session 1. For the second half of session 2, you will no longer be given any feedback (neither positive or negative feedback). Your bonus points will appear as X's (XXXX) onscreen, and every response, whether correct or incorrect will be confirmed by an "indifferent" face. You will still receive bonus points, so continue to try diligently despite the lack of feedback.

- Although each of the four categories looks very similar, I can tell you several important properties that distinguish them from one another. First, all images are composed of a jagged-edged black circle with three spots inside. The spots are the critical component that determines what category an image belongs to. In particular, the middle spot is very important.

- Two things I should let you know about the experimental room: we're going to have you place your chin on a chin rest during the experiment. This is just to make sure that your position relative to the screen remains the same. Also, the experimental room isnt very large, just to give you a heads up in case you might feel uncomfortable in a small room.

- Again, there are no known risks associated with this experiment, and all data is stripped of any identifying characteristics.
- Do you have any questions about participation?