(2011). In A. Cangelosi, J. Triesch, I. Fasel, K. Rohlfing, F. Nori, P.-Y. Oudeyer, M. Schlesinger, Y. and Nagai (Eds.), *Proceedings of the First Joint IEEE Conference on Development and Learning and on Epigenetic Robotics*. New York: IEEE.

Increasing Spatial Competition Enhances Visual Prediction Learning

Matthew Schlesinger
Brain and Cognitive Sciences
Department of Psychology
Southern Illinois University
Carbondale, IL 62903
Email: matthews@siu.edu

Dima Amso
Cognitive, Linguistic, and
Psychological Sciences
Brown University
Providence, RI 02912
Email: Dima_Amso@brown.edu

Scott P. Johnson
Department of Psychology
University of California, Los Angeles
Los Angeles, CA 90095
Email: scott.johnson@ucla.edu

***Abstract*—Our previous work provides support for the idea that the development of visual perception in early infancy depends on progressive improvements in oculomotor skill. In particular, we have proposed and tested an eye-movement model that successfully reproduces infants' gaze patterns on two measures of visual attention. However, this result is due to explicit hand-tuning of a key parameter in the model. In the current simulation study, we investigate whether manipulating this parameter (i.e., the duration of spatial competition) enhances visual prediction learning. As expected, we find that prediction learning becomes more accurate as spatial competition in the eye-movement model is increased. This finding suggests that visual prediction learning can provide a meaningful error-feedback signal, which can be used to modulate spatial competition in the eye-movement model.**

***Index Terms*—perceptual development, spatial competition, visual prediction learning, oculomotor skill.**

I. Introduction

As a biological adaptation, foveal vision is a "double-edged sword." On the one hand, sampling a small portion of the visual field at a high spatial resolution helps to reduce the visual-information bottleneck. On the other hand, successful information pick-up depends on effective and efficient scanning of the environment, and this capacity is not an innate ability in human infants, but is instead a learned skill [1].

Oculomotor skill develops rapidly after birth, supported in part by maturation and growth of the cortical processing areas that are associated with vision [2]. Across a series of studies, using both conventional behavioral methods with young infants as well as simulations of infants' gaze patterns, we have focused on three core questions [3-6]. First, what is the pattern of oculomotor skill development in early infancy? Second, what biological and environmental factors make this developmental pattern possible? Third, how do improvements in oculomotor skill create the opportunity for infants to discover critical visual features in their environment?

One of the ways that we have investigated these questions is with the **perceptual-completion task**, which is presented in Figure 1. In this paradigm, infants first view the display illustrated in Figure 1A, in which a green rod moves laterally behind a blue screen. As we describe in the next section, infants' subsequent responses to the events displayed in Figures 1B and Figure 1C provide a means for assessing whether they perceive the occluded rod as a coherent, solid object (we refer to this phenomenon as *unity perception*) or if instead they view it as two disjoint surfaces moving at the same time.

Our behavioral and simulation results to date suggest that there is a fundamental connection between how infants distribute their attention over the visual scene and how they develop the capacity for unity perception. In particular, we have developed and tested a computational model that simulates infants' eye movements as they view displays like the one in Figure 1. A key parameter in the model, which controls the duration of "competition" among potential fixation targets, successfully accounts for infants' performance not only on the perceptual completion task, but also on a visual search task.

However, this developmental pattern does not emerge in the model autonomously, but is instead generated by hand-tuning of the spatial-competition parameter. The goal of the current simulation study, therefore, is to identify an appropriate performance metric that can be used as a training signal within the model to adaptively modulate this parameter.

The metric that we have chosen to investigate is *visual prediction learning*. In Section II, we motivate this choice by

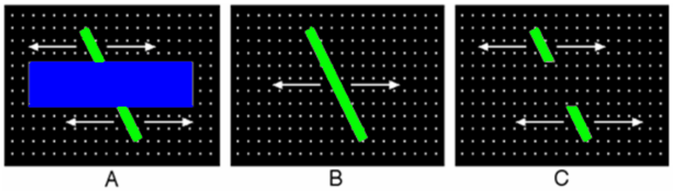

Fig. 1 Displays used to assess perceptual completion in infants: (A) habituation display, and (B) complete rod and (c) broken rod test displays.

first describing how infants' gaze patterns during the occluded-rod display are correlated with their performance on the perceptual-completion task. Next, Section III provides an overview of the eye-movement model, including the neurophysiological pathways and processes that it is designed to represent. In this section, we focus on the mechanism of spatial competition, and propose the hypothesis that increasing spatial competition will lead to an enhancement in visual prediction learning. Next, the results of the simulation study are presented in Section IV, while in the final section, we discuss potentials ways to exploit error-feedback from the prediction-learning system as a training signal in the eye-movement model.

II. The Development of Perceptual Completion

Perceptual completion is assessed by first presenting infants with the occluded-rod display until they habituate. The two post-habituation test events (Figures 1B and 1C) are then presented and infants' looking times to each of the displays are compared. In particular, longer looking to one of the two test displays is assumed to reflect a novelty preference. Therefore, infants who perceive the occluded rod as a coherent, unified object (i.e., *unity perception*) are assumed to experience the complete rod as a familiar display, and therefore show a preference for the broken-rod display. These infants are referred to as *perceivers*. Alternatively, infants who experience the occluded rod as two disjoint or disconnected surfaces will look longer at the complete-rod display, and are therefore referred to as *nonperceivers*.

As a behavioral measure of occluded-object perception, the perceptual-completion task highlights the first four months of postnatal life as an important time period for human infants [3-4, 7-8]. Between birth and age 2 months, infants are typically classified as nonperceivers, as they look longer at the complete-rod test display (Figure 1B) [7-8]. By age 4 months, however, unity perception is a relatively robust phenomenon: 4-month-olds look reliably longer at the broken-rod test display (Figure 1C), indicating that they perceive the occluded rod as a single, unified object [8].

How does the development of oculomotor skill influence the transition from nonperceiver to perceiver, between ages 2 and 4 months? This question was investigated by [3], who first measured 3-month-olds' eye movements during the occluded-rod display. Infants then viewed the posthabituation test displays, and were classified as either perceivers or nonperceivers.

Figure 2 presents representative scanplots produced by two infants during the occluded-rod display [3]. Note that the infant on the left, who was categorized as a nonperceiver, distributed a large portion of their fixations toward the occluding screen. In contrast, the infant on the right, who was categorized as a perceiver, produced more fixations toward the moving, occluded rod. Each of these qualitative patterns is not only typical of nonperceivers and perceivers, respectively, but in a follow-up study, [4] also found that perceivers generated a significantly higher proportion of rod scans (i.e., fixations toward the rod segments) than nonperceivers. Taken together, these findings highlight the role of oculomotor skill, and in particular, they demonstrate that perceivers produce gaze patterns that are more effectively distributed over the occluded-rod display.

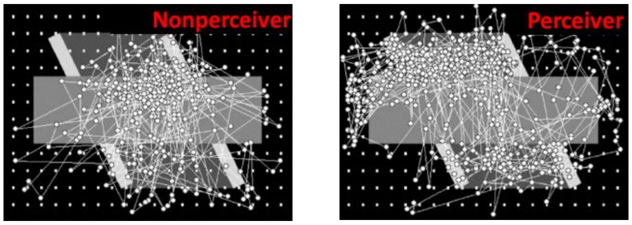

Fig. 2 Sample scanplots from two infants studied by [3]. Nonperceivers tend to view less-relevant areas of the display (e.g., occluding screen) while perceivers spend more time viewing the moving, occluded rod.

III. The Eye-Movement Model

Our subsequent work has focused on designing and testing an eye-movement model that captures the developmental transition from nonperceivers to perceivers, while identifying relevant brain areas that may provide a substrate for visual attention and serve as a mechanism for developmental change [5-6].

A. *Overview of the eye-movement model*

The model is an extension of the salience-map framework proposed by [9-10]. Processing within the model occurs over four stages, three of which are illustrated in Figure 3:

1) *Retinal Image (3A).* An input image is projected onto the model's simulated retina.

2) *Feature Maps (3B).* The retinal image projects through four optical filters (i.e., intensity, motion, color, and oriented edges), which decompose the input image into a set of retinotopic feature maps. During this stage, a spatial-competition filter is applied to each feature map.

3) *Salience Map (3C).* The feature maps are pooled into a single salience map.

4) *Eye Movement (not illustrated).* A highly-active location on the salience map is selected probabilistically, and an eye movement to this location is produced.

Across a series of simulation studies, we have examined a number of parameters in the model that each represent a specific brain region or neural mechanism. We focus here on one parameter in particular, which is an analog for the duration of recurrent activity in posterior parietal cortex, a region of the brain that is associated with the encoding of visual salience and the modulation of visual attention [11-12].

This parameter of interest is utilized during the final step in the feature map process (see Figure 3B), in which a spatial-competition filter is applied to each of the feature maps. At a computational level, this filter distributes activity at each location on the feature map outward to a local neighborhood,

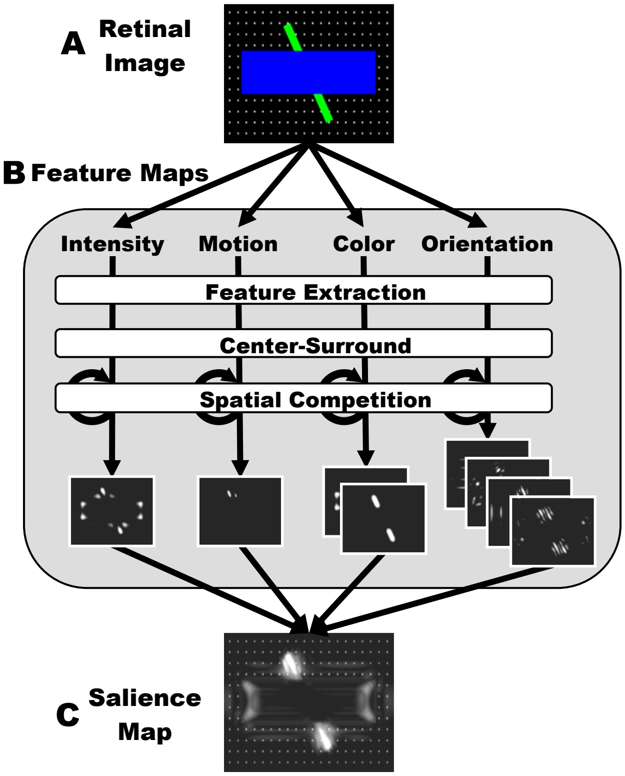

Fig. 3 Schematic diagram of the eye-movement model: (A) An input image is projected onto the retina, (B) the retinal image is projected through four feature channels (intensity, motion, color, and orientation), and (C) feature maps produced across the four feature channels are pooled into a single, unified salience map.

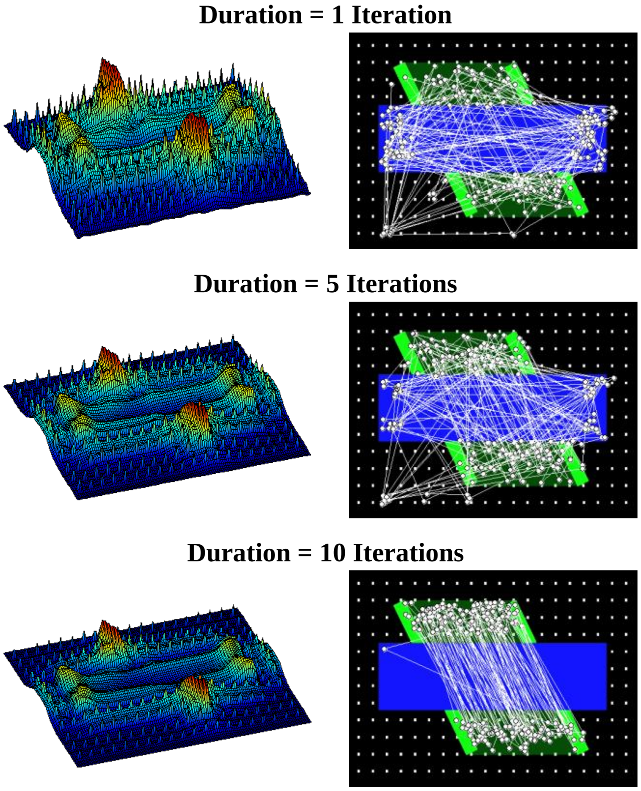

Fig. 4 Three examples of the eye-movement model's salience map and corresponding scanplots. Left panel: snapshots of the salience map during the perceptual-completion task, with spatial competiton set at 1, 5, and 10 iterations. Right panel: cumulative scanplots over a series of trials, for the same respective values of spatial competition.

while also inhibiting activity across the map globally.

Note that the spatial-competition filter is represented within the model as a discrete, iterative process, which can be applied an arbitrary number of times prior to the third stage (i.e., production of the salience map). Our work to date demonstrates that by hand-tuning this parameter through a range of possible durations (i.e., number of iterations), the model not only captures infants' performance on the perceptual completion task, but also the visual search task investigated by [4].

How does the duration of spatial competition have an effect on the model's gaze patterns? Figure 4 helps to address this question, by illustrating how variation in the spatial-competition parameter influences the topology of the salience map, which in turn determines the model's gaze patterns during visual input. The left panel of Figure 4 presents a snapshot of the salience map during the perceptual-completion task, after 1, 5, and 10 iterations of the spatial-competition filter. The right panel presents cumulative scanplots of the model's fixations as a function of the three durations of spatial competition.

After 1 iteration of spatial competition, the salience map includes not only activation peaks associated with the rod segments and the left and right edges of the occluder, but also a number of smaller peaks that are due to the background and the edges of the occluder. By the 5th iteration, however, many of the smaller activation peaks have diminished, and by the 10th iteration most are no longer present.

This pattern is reflected in the corresponding scanplots. In particular, when the duration of spatial competition is set at 1 iteration, 7% of the fixations are generated toward the rod segments. This proportion increases to 28% with 5 iterations, and up to 81% with 10 iterations. As a point of reference, in [6] we found that the eye-movement model matches the proportion of rod scans produced by nonperceivers when the duration of spatial competition is 3 iterations, and that it matches perceivers when the duration is 4 iterations.

B. Predicton learning as an error signal

An important limitation of our model is that the spatial-competition parameter is hand-tuned. One way to address this issue is to introduce a link between the gaze-control and feature map components of the model. In particular, prediction errors in the gaze-control system can be used as a feedback or "training signal" that tunes duration of spatial competition within the feature map system [13-14].

One cortical region that might provide such feedback is the frontal eye fields (FEF), which is part of the premotor area and is associated with voluntary eye movements. Recent findings highlight two key functional properties of FEF activity. First, stimulation of FEF neurons during tracking of occluded

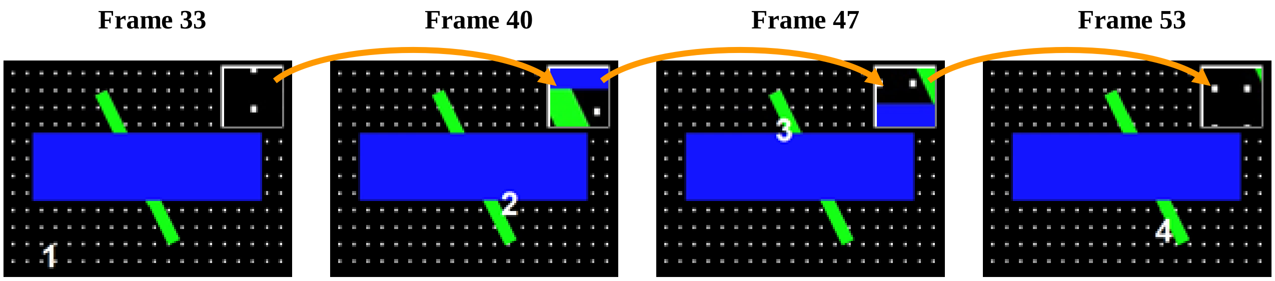

Fig. 5 Illustration of the sampling procedure used to produce a sequence of foveal samples during training of the prediction model (normal-order condition). A sequence of four fixations are presented (1-4), and the corresponding foveal samples are illustrated in the upper-right corner of each image. Orange arrows indicate that the task of the prediction model is to learn the foveal sample that will result from the next eye movement.

objects results in anticipatory eye movements [15]. Second, and more importantly, FEF neurons are capable of producing an error signal that corrects an ongoing eye movement [16]. Together, these findings implicate the FEF as a brain area that not only computes future sensory states, but that can also generate an error-feedback signal during visual activity.

Before investigating this idea, we decided to first address a more basic question: how will changes in the level of spatial competition within the model influence visual prediction learning? In order to address this question, we introduced a new component to the model, which simulates the process of visual prediction learning.

In particular, as a proxy for prediction learning in the FEF we used a 3-layer feedforward neural network, which is provided with a "foveal sample" centered on the eye-movement's current gaze location, and is trained to predict the next foveal sample (in other words, a forward model). Figure 5 illustrates the idea: imagine that on frame 33 of the occluded-rod display, the current gaze location is the lower left corner of the display (Figure 5, #1). The foveal sample associated with this location is illustrated in the upper right corner of the image. The task of the prediction network, given this foveal sample, is to predict the next foveal sample that will occur after the subsequent eye movement (i.e., approximately 210ms later, on frame 40).

We next describe in greater detail how this process was generalized as a training procedure, in order to investigate the question of whether modulation of the spatial-competition parameter would result in improved prediction learning. Given the qualitative findings illustrated in Figure 4—that is, that increasing spatial competition appears to increase "clustering" of the fixations—we proposed two specific hypotheses:

1) *Spread of fixations.* Increasing spatial competition will result in less-disperse fixations over the occluded-rod display.

2) *Enhanced visual prediction learning.* Due to less disperse fixations, increasing spatial competition will also result in more accurate visual prediction learning.

C. Training the prediction model

In order to train the prediction model, we first used the eye-movement model to produce 11 sets of gaze sequences generated in response to the occluded-rod display. Specifically, each set corresponded to first setting the spatial-competition parameter to a value in the range [0, 11]. Next, the occluded-rod display was presented to the model over 10 repetitions. Each repetition lasted 5 seconds, and resulted in approximately 22 fixations (additional details of the simulation process are presented in [6]). This process constituted a single simulated *trial*, and was repeated 10 times per parameter value, resulting in 10 trials and approximately 220 fixations (i.e., foveal samples) per spatial competition value.

Next, the prediction model was trained offline under three training regimes. In each case, training began by first obtaining a foveal sample from the occluded-rod display (see Figure 5). The size of the sample was a 41x41-pixel square, centered on the corresponding fixation point. As the entire display was 480x320 pixels, the foveal sample was roughly 1% of the full display.

Each sample was transformed from RGB to grayscale representation, and fed into a 3-layer network with 1681 input units, 400 hidden units (i.e., 25% fan-in), and 1681 output units. The subsequent foveal sample was set as the target output, and differences between observed and target output were used to update connections via the backprop learning algorithm.

In order to systematically examine the influence of modulating spatial competition, three training conditions were compared:

1) *Normal order.* For each duration of spatial competition, the original (i.e., canonical) order of fixations was used to train the prediction model.

2) *Mixed order.* For each duration of spatial competition, the original fixations were put in random order. This allowed us to tease apart the influence of fixation spread or dispersion, independent of the temporal order of the fixations.

3) *Random order.* As a baseline condition, the model was trained on fixations that were selected from the occluded-rod display at random.

A complete pass through one set of gaze sequences (i.e., 10 trials) constituted an *epoch*. For each duration of spatial competition, the model was trained for 500 epochs, which defined a training *run*. Note that for all three training

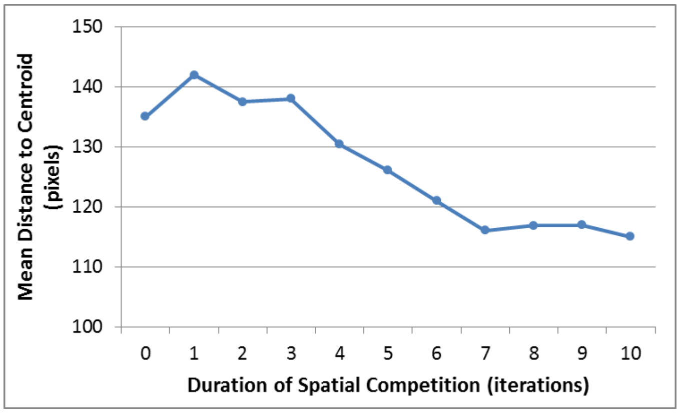

Fig. 6 Spatial dispersion of fixations produced by the eye-movement model, as a function of the duration of spatial competition. Dispersion was measured as the mean distance from the centroid, within each set of gaze sequences.

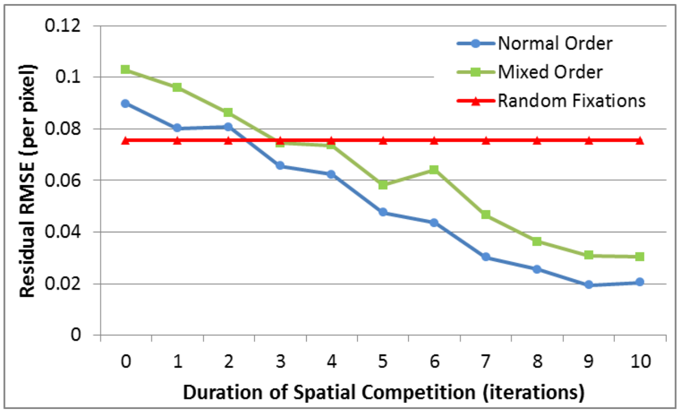

Fig. 7 Residual mean squared error (RMSE) in the prediction model, averaged over the final 20 epochs of training. The three conditions tested were: normal order (blue), mixed order (green) and random fixations (red).

conditions (including the random-order condition), within a run, gaze sequences were held constant across epochs. Connections in the network were randomly initialized at the start of each run, and 10 training runs were performed for each of the 11 durations of spatial competition.

IV. Results

We focus here on two analyses. In the first analysis, we examine our hypothesis that fixations should become less disperse as the duration of spatial competition is increased. Next, the second analysis evaluates our second hypothesis, that is, that increasing the duration of spatial competition will result in more accurate prediction learning.

A. Analysis 1: Spread of fixations

Our informal analysis of the salience maps suggested that as spatial competition increases, the number of peaks on the maps decreases. As Figure 4 illustrates, this should decrease the spread of fixations, as these are selected probabilistically as a function of activation-level on the salience map (i.e., higher-active areas are more likely to be selected for fixation).

The first analysis sought to confirm this observation, by systematically measuring the dispersion of fixations for each of the 11 gaze sequences. We therefore computed the 2D centroid for each gaze sequence, and then calculated the mean distance of the fixation points within each sequence from the corresponding centroid (a comparable result is obtained by computing the standard deviation rather than the mean). Figure 6 presents the results of this analysis: as expected, fixations became less disperse as the duration of spatial competition was increased (excluding increases from 0 to 1, and 2 to 3 iterations). It is interesting to note that the largest decrease occurs between 3 and 4 iterations; these, perhaps coincidentally, are the same values at which the model captures the performance of nonperceivers and perceivers, respectively, on the perceptual-completion task.

B. Analysis 2: Visual prediction learning

We next analyzed the performance of the prediction model, as a function of the duration of spatial competition. Prediction error was defined as the root mean squared error (RMSE) produced by the network, averaged over the number of pixels in the output layer (i.e., 1681). Figure 7 presents the "residual" RMSE, which is the mean error per pixel in the model, averaged over the final 20 epochs of training.

We highlight here three important results:

1) *Normal-Order Condition.* First, as expected, increasing the duration of spatial competition enhanced visual prediction learning. Specifically, residual prediction errors decreased in the model as the duration of spatial competition was increased (Figure 7, blue line).

2) *Mixed-Order Condition.* Next, we found that the order of fixations produced by the eye-movement model also had an effect on prediction learning. In particular, when the original fixations were learned in an arbitrary (i.e., mixed) order, the residual RMSE increased for all 11 durations of spatial competition (Figure 7, green line). This suggests a second, temporal influence on prediction learning, independent of the spatial spread of fixations.

3) Random-Fixations Condition. Finally, Figure 7 also plots residual RMSE in the random-fixations condition (red line). Since training in this condition did not depend on the spatial competition parameter, the same result is plotted at a constant level. In general, prediction learning was more successful in the majority of the normal-order and mixed-order conditions. However, it is worth noting that between 0 and 3 iterations of spatial competition, the prediction model actually learned to predict gaze sequences that were generated at random more accurately then either the normal- or mixed-order sequences.

V. Conclusion

The results were consistent with each of our hypotheses. First, increasing spatial competition results in a gradual transformation of the salience map from a landscape initially populated with numerous potential fixation targets, to one with only a few strong areas of activity. Because fixations are selected as a function of salience, we hypothesized that increasing spatial competition would result in a less-disperse distribution of fixations. This prediction was confirmed in the first analysis.

Next, we also hypothesized that increasing the duration of spatial competition should enhance visual prediction learning. This is a relatively straightforward hypothesis, given the intuition that greater clustering of fixations should also result in less variation among foveal samples, or in other words, a smaller training set with more similar input patterns. Indeed, the findings also confirmed this hypothesis.

However, two additional findings suggest that reducing the spread of fixations does not completely account for the improvement in visual prediction learning. First, note that when the prediction model was trained on the original gaze sequences in a scrambled order, performance declined. Therefore the specific temporal order in which the fixations occur provides an additional regularity or statistical cue that the prediction model is able to detect and exploit.

Second, note that while increasing the duration of spatial competition enhances visual prediction learning in general, we also found at the lowest levels of spatial competition (i.e., between 0 and 3 iterations) that the prediction model actually performed better when it was trained on a random sequence of fixations. This finding provides an important clue about the relation between oculomotor skill and visual prediction learning: in particular, it suggests that a minimum level of oculomotor skill may be necessary before visual prediction learning can be used effectively (i.e., in some cases, random input is better than non-random input).

Recall that our primary goal was to determine whether performance on the visual prediction learning task could be used to "bootstrap" adaptive tuning of the spatial competition parameter in our eye-movement model. Taken together, the results of the current simulation study provide a very strong, positive answer to this question.

Our current work is pursuing this issue in two parallel directions. First, we are developing a hybrid model that allows the prediction and eye-movement systems to share information. In particular, we are using a variant of the actor-critic architecture, in which errors in the prediction model serve as an error feedback signal that alters the duration of spatial competition in the eye-movement model, via reinforcement learning.

Second, we are also expanding our testing of the eye-movement model to a library of natural images, in order to determine whether the effect of increasing spatial competition on visual prediction learning generalizes to stimuli beyond the perceptual-completion task.

REFERENCES

[1] Haith, M.M., *Rules that babies look by: The organization of newborn visual activity*. New Jersey: Erlbaum, 1980.

[2] M.H. Johnson, "Cortical maturation and the development of visual attention in infancy," *Journal of Cognitive Neuroscience*, vol. 2, pp. 81–95, 1990.

[3] S.P. Johnson, J.A. Slemmer, and D. Amso, "Where infants look determines how they see: Eye movements and object perception performance in 3-month-olds," *Infancy*, vol. 6, pp. 185-201, 2004.

[4] D. Amso, and S.P. Johnson, "Learning by selection: Visual search and object perception in young infants," *Developmental Psychology*, vol. 42, pp. 1236-1245, 2006.

[5] M. Schlesinger, D. Amso, and S.P. Johnson, "The neural basis for visual selective attention in young infants: A computational account," *Adaptive Behavior*, vol. 15, pp. 135-148, 2007.

[6] M. Schlesinger, D. Amso, and S.P. Johnson, "Simulating infants' gaze patterns during the development of perceptual completion," *Proceedings of the Seventh International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, vol. 7, C.G. Prince, M. Littman, H. Kozima, and C. Balkenius, Eds. Sweden: Lund University Cognitive Studies, pp. 157-164, 2007.

[7] A. Slater, S.P. Johnson, E. Brown, and M. Badenoch, "Newborn infants' perception of partly occluded objects," *Infant Behavior and Development*, vol. 19, pp. 145-148, 1996.

[8] S.P. Johnson, "Development of perceptual completion in infancy," *Psychological Science*, vol. 15, pp. 769-775, 2004.

[9] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, pp. 1489-1506, 2000.

[10] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual-attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254–1259, 1998.

[11] J.P. Gottlieb, M. Kusunoki, and M.E. Goldberg, "The representation of visual salience in monkey parietal cortex," *Nature*, vol. 391, pp 481-484, 1998.

[12] K.M. Shafritz, J.C. Gore, and R. Marois, "The role of the parietal cortex in visual feature binding," *Proceedings of the National Academy of Scienc*es, vol. 99, pp. 10917-10922, 2002.

[13] C., Balkenius, and B. Johansson, "Anticipatory models in gaze control: A developmental model," *Cognitive Processing*, vol. 8, pp. 167-174, 2007.

[14] C. Weber, and J. Triesch,. "A possible representation of reward in the learning of saccades," Proc*eedings of the Sixth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems* , vol. 6, F. Kaplan, P.Y. Oudeyer, P. Gaussier, J. Nadel, L. Berthouze, H. Kozima, C. Prince, and C. Balkenius, Eds. Sweden: Lund University Cognitive Studies, pp. 153-160, 2006.

[15] A. Barborica, and V.P. Ferrera, "Modification of saccades evoked by stimulation of frontal eye field during invisible target tracking," *Journal of Neuroscience*, vol. 24, pp. 3260-3267, 2004.

[16] V.P. Ferrera, and A. Barborica, "Internally generated error signals in monkey frontal eye field during an inferred motion task," *Journal of Neuroscience*, vol. 30, pp. 11612-11623, 2010.