Towards a What-and-Where Model of Infants' Object Representations

Matthew Schlesinger and Roberto Limongi

Brain and Cognitive Sciences Program
Department of Psychology
Southern Illinois University
Carbondale, IL 62901
matthews@siu.edu robertol@siu.edu

Abstract

We propose that the capacity for infants to form mental representations of hidden or occluded objects can be decomposed into two tasks: one process that identifies salient objects, and a second complementary process that identifies salient locations. This approach is inspired by the distinction between dorsal and ventral extrastriate visual processing in the primate visual system. We illustrate our approach by describing an eye-movement model that incorporates both dorsal and ventral processing streams, and by using the model to simulate infants' reactions to possible and impossible events from an infant looking-time study (Baillargeon, 1986). As expected, we find that the dorsal system is sensitive to the location of a key feature in these events (i.e., the location of an obstacle), while the ventral system responds equivalently to the possible and impossible events. We conclude by discussing how these results may help explain infants' reactions in looking-time studies.

Introduction

As we interact with the physical world (e.g., climb stairs, pick up a coffee mug, swing a tennis racket, etc.), our actions imply knowledge of solid, 3-dimensional objects that obey the principles of Newtonian mechanics. This implicit knowledge is so compelling and intuitive that developmental researchers have begun to suggest that it may in fact be part of our genetic heritage (e.g., Baillargeon, 1999; Spelke, 1998; however, see also Haith, 1998; Smith, 1999).

Much of the research on infants' object knowledge (e.g., Baillargeon, 1995) has focused on how infants are able to form mental representations of occluded or hidden objects, which not only provide an internal cue or symbol for the continued existence of real-world objects (i.e., object permanence), but also for the their physical properties (e.g., location, shape, color, motion path, etc.).

For developmental robotics, this work raises three fundamental challenges: (1) to identify and describe the object representations used by infants, (2) to translate these representations into a computationally explicit form (i.e., implement them in an algorithm), and (3) to design a computational model that successfully incorporates and exploits these representations.

In this paper, we present a research strategy for addressing these challenges that is inspired by both structural and functional properties of the primate visual system. Specifically, we describe how the extrastriate dorsal and ventral streams can be used to develop a neural network model of infants' object representations, in which objects and their various properties are encoded along two parallel pathways.

The remainder of the paper is organized as follows. In the next section, we review the properties of the dorsal and ventral pathways. We then briefly describe the "car study" (Baillargeon, 1986; Schlesinger & Casey, 2003), which we use as a platform for developing and testing our model. Next, we present the key features of our eye-movement model, and describe how it is used to simulate infants' gaze patterns in the car study. We conclude by describing the performance of the model, and discussing some of the future directions of this work.

What-and-Where as Task Decomposition

The key reason we focus on the dorsal-ventral distinction as a modeling strategy is that it provides a "natural" decomposition of the visual world into two distinct types of visual representations. First, the dorsal or "where" pathway travels from occipital to parietal cortex, and is functionally specialized for spatial processing, and in particular, spatially-oriented action such as reaching and visual tracking (e.g., Milner & Goodale, 1995). Two key features of dorsal processing are a high sensitivity to contrast, as well as to motion (particularly in the periphery; see Steward, 2000).

Second, the ventral or "what" pathway travels from occipital to temporal cortex, and is functionally specialized for visual form analysis and object processing such as face recognition (e.g., Mishkin, Ungerleider, & Macko, 1985).

Consequently, ventral processing relies on high-resolution information (particularly from the fovea; see Steward, 2000).

In addition to neurophysiological evidence that supports this distinction (e.g., patients with focal lesions in either parietal or temporal cortex), computational studies also provide support for the idea that finding and recognizing objects is more efficiently accomplished by decomposing the problem into two parallel tasks—(1) identifying objects versus (2) localizing objects—compared to solving both tasks with only one system (e.g., Jacobs, Jordan, & Barto, 1991; Rueckl, Cave, & Kosslyn, 1989).

How do these two systems develop in human infants? A general developmental pattern found across a wide variety of visual processing tasks is that the dorsal pathway develops first, while the ventral pathway appears to develop more slowly over the first year (e.g., Leslie, Xu, Tremoulet, & Scholl, 1998; Mareschal, Plunkett, & Harris. 1999). In addition, the dorsal and ventral streams may not start to become coordinated and integrated until the end of the first year (Mareschal & Johnson, 2003).

The dorsal-ventral distinction also offers an important insight for the study of infants' early object representations. As we highlight in the next section, research in this area often assumes that infants have the capacity for representation, while failing to provide a detailed account of the cognitive mechanisms that make representation possible. Therefore, a dorsal-ventral model not only suggests an explicit mechanism for representing salient objects and their locations, but also is based on known principles from developmental visual neuroscience.

The "Car Study"

The car study was designed and first investigated by Baillargeon (1986; Baillargeon & DeVos, 1991). In this study, infants watch a simple mechanical display, in which a car rolls down a ramp, behind a screen, and out the other side. Figure 1A presents a schematic display of this **familiarization** event. Note that at the start of the familiarization event, the screen is raised in order to show the infant that nothing is behind it.

After watching several repetitions of the familiarization event, infants then see two novel test events (see Figures 1B and 1C). During both the **possible** and **impossible** test events, a box is revealed behind the screen. During the impossible event, however, the box is placed on the track, in the path of the car. Nevertheless, during both test events the car reappears after passing behind the screen.

Baillargeon found that by age 6 months infants look significantly longer at the impossible event than the possible event. How did she interpret these findings? First, she suggested that infants mentally represent both the occluded box and the car as it passes behind the screen. Second, she proposed that infants use these representations to "compute" when the car should reappear, and are consequently surprised to see the car reappear during the impossible event, when its path is obstructed by the box. Thus, because the impossible event is surprising or unexpected to infants, they spend more time looking at it.

While informative, these findings unfortunately do not address two crucial questions: *how* do infants represent occluded objects such as the box and car, and *what* are the nature of these representations (i.e., what features are encoded and stored)? This is due, at least in part, to the fact that looking-time studies with infants rely on overt behaviors as an index for their expectations, which are not directly observable.

The dorsal-ventral distinction, in contrast, offers a way to deal with both of these questions. First, it addresses the *how* of representation by suggesting that perceptual data is

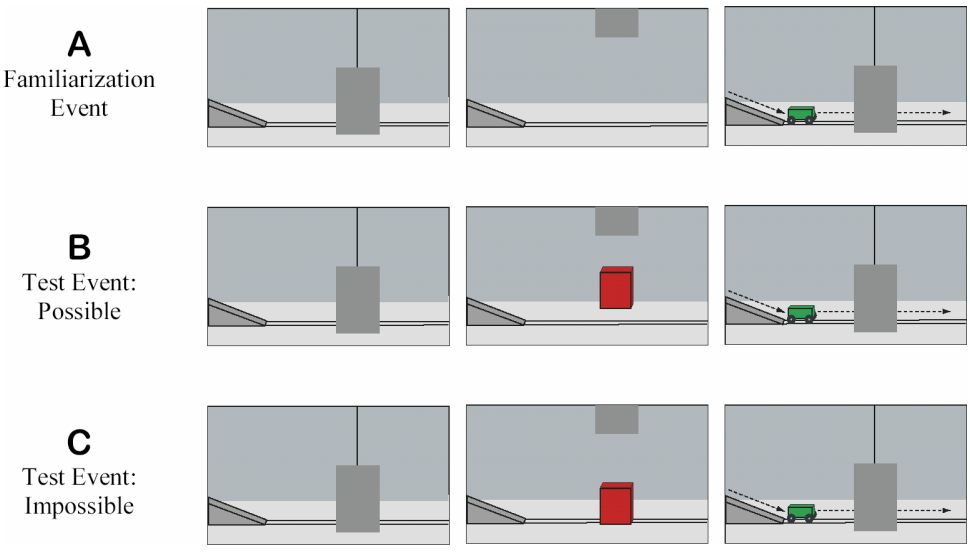

Figure 1: Schematic display of the familiarization (A), possible (B), and impossible (C) events in the car study.

maintained along the dorsal and ventral paths via persistent neural activity patterns, which provide a memory trace during breaks in contact with an object (e.g., occlusion). Second, it also addresses the *what* of representation by suggesting that visual scenes are decomposed into at least two fundamental categories, salient objects on the one hand, and salient locations on the other.

We pursued this argument by reasoning that because the *location of the box* is a critical feature in the car study, it should be the *dorsal pathway* that is primarily responsible for detecting this feature during the possible and impossible events, and consequently, drawing attention to the relevant location (i.e., either behind or on the track). In particular, we proposed two key predictions. First, we predicted that since there are no changes in the appearance of the box during the test events, there should be no significant differences in the response of the ventral system to the possible and impossible events (i.e., both events should be equally "novel"). Second, and in contrast, we predicted a significant difference in the response of the dorsal system during the test events. Specifically, we predicted that the dorsal system would show a stronger "novelty" reaction to the impossible test event.

Modeling Infants' Gaze Patterns

The present model is based on a platform that we have developed to simulate infants' gaze patterns in looking-time studies (Schlesinger, 2003; Schlesinger & Barto, 1999; Schlesinger & Parisi, 2001; Schlesinger & Young, 2003). Key elements of the model include: (1) a simulated retina, with both low- and high-resolution input (analogous to the periphery and fovea), (2) an artificial neural network that serves as an oculomotor control system, and (3) simulated eye-movements that enable the fixation point to move over time.

There are two major innovations in the present model. First, we used digitized video (rather than animation) as input to the model, which was recorded in our lab from the same location where infants sit as they view the car study (Schlesinger & Casey, 2003). Second, as we highlight below, the architecture of the model has been elaborated to include three processing streams or pathways (i.e., the dorsal, ventral, and also the superior colliculus pathways). While our approach is similar to a recent model that also simulates the dorsal and ventral pathways in infants (Mareschal, Plunkett, & Harris, 1999), three unique features are: (1) the use of digital video input, (2) production of eye movements, and (3) two levels of visual resolution.

Model Architecture

Figure 2A presents a schematic diagram of the eye-movement model, including the major processing pathways leading from visual input to eye movements. Note that gray boxes in the diagram represent processing stages (white boxes are "pass through"), and that boxes with dotted borders have modifiable parameters (see sections 3 and 4, below). Figure 2B illustrates activity along two of the pathways (i.e., dorsal and ventral) during a sample input frame, after 300 training trials (see **Training and Testing**, below).

1. Input. Three events from the car study (i.e., familiarization, possible, and impossible; see Figures 1A, 1B, and 1C, respectively) were produced in our lab and recorded with a digital video camera at the rate of 30 frames per second (for details on the design of the apparatus, see Schlesinger & Casey, 2003). The duration of each event was 7 seconds, and each frame was 240 by 180 pixels (in grayscale). The events were then parsed into image sequences, for a total of 210 image frames per event.

All frames were pre-processed prior to training. In particular, low-resolution images were obtained by reducing each frame to 20% of its original size (i.e., 48 by 36 pixels). Similarly, each low-resolution frame was also preprocessed with motion and edge filters.

2. Superior Colliculus Path. The superior colliculus is part of the retinotectal pathway, and represents a functionally "older" part of the mammalian brain that is devoted to motion processing (Steward, 2000). We included this path as it appears to provide a basic cue for motion to infants soon after birth, and may function as a bootstrap that complements motion processing in cortical regions (e.g., area MT). Consequently, as Figure 2A indicates, low-resolution motion frames pass through the superior colliculus toward the saliencemap. Motion is computed by taking the absolute value of the difference between consecutive frames, and setting all non-zero pixel values (i.e., those that change between frames) to 1.

3. Dorsal Path. To capture the key roles of contrast and motion processing in the dorsal pathway, we combined low-resolution edge and motion frames into a single image, and used these as input into the dorsal system. Because the dorsal path plays a critical role in spatially-oriented actions, we implemented a prediction-learning algorithm as a proxy for action-guidance, in which the task of the dorsal system is to learn to predict the next image frame for each input frame that it receives (for a related approach, see Schlesinger & Young, 2003). In particular, we employed a 3-layer, fully-connected network (1728, 172, and 1728 units, respectively).

The left side of Figure 2B illustrates processing along the dorsal pathway during a sample input frame (the red "x" indicates the current fixation point). In particular, while the car is visible in the input frame, the dorsal system fails to correctly reproduce it in the output. In this case, note that dorsal "error" (the absolute value of dorsal output minus dorsal input) is maximal in the region of the car.

4. Ventral Path. Because the ventral path relies on high-resolution visual input, we sampled a 30-by-30 pixel region from each high-resolution frame, centered on the current fixation point during that frame, as input to the

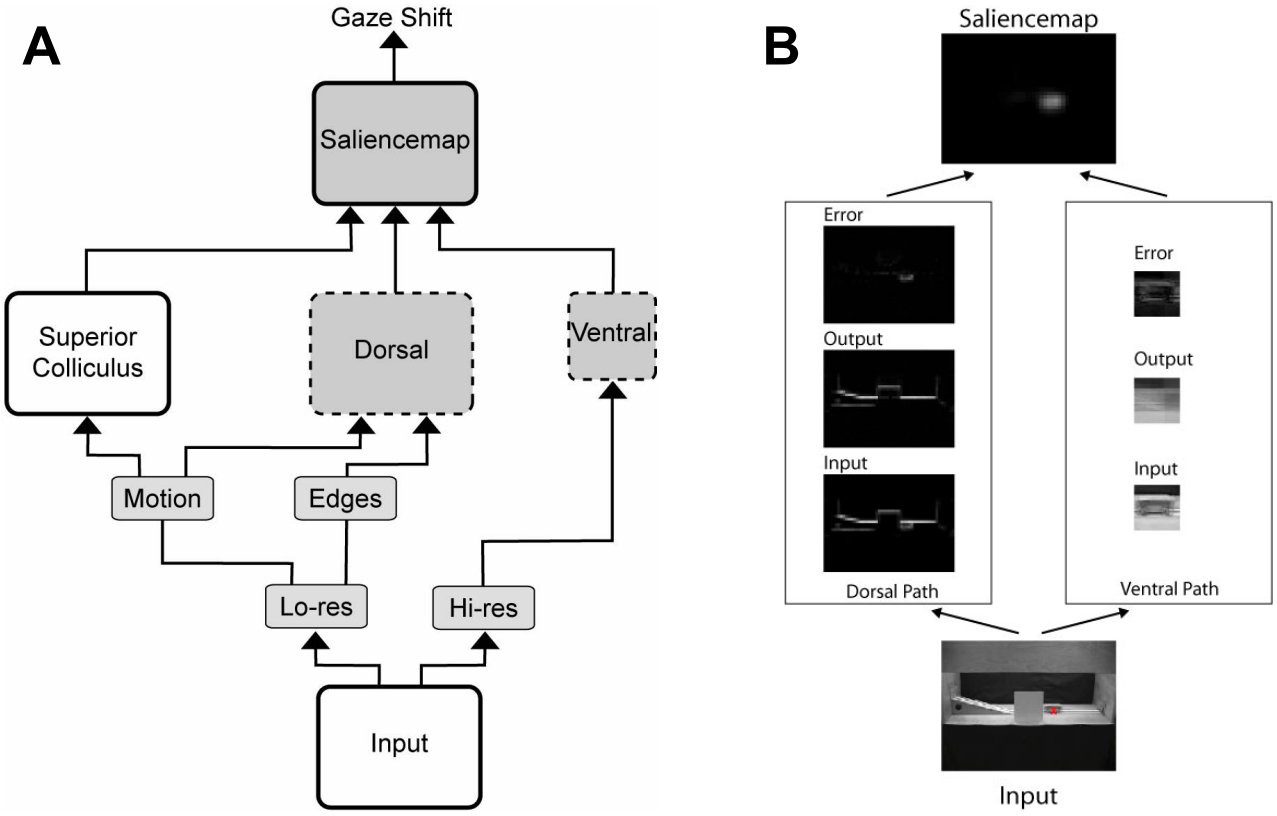

Figure 2: Schematic diagram of the eye-movement model, including (A) model architecture and (B) processing of a sample input frame through the dorsal and ventral systems after 300 training trials (superior colliculus path not shown).

ventral system. Specifically, the ventral system was implemented as a 3-layer, fully-connected autoencoder network (990, 90, and 900 units, respectively), including recurrent connections from the hidden layer back to the input layer. The task of this system is to learn to reproduce the input over a set of output units. Consequently, encoding "errors" were used as a proxy for recognition errors in the ventral system.

As the right side of Figure 2B illustrates, the ventral system also has some difficulty in processing the car after 300 training trials. In particular, while the ventral network "recognizes" (i.e., reproduces on the output units) the general shades of the foreground and background, the details of the car are missing from the output, and therefore, are associated with regions of error in the output.

5. Saliencemap. The last stage of processing in the model is the saliencemap, which combines information from the earlier pathways into a single, coherent representation (for a related approach, see Itti & Koch, 2000).

Specifically, the saliencemap sums input from the superior colliculus system (i.e., a binary motion map), the dorsal system (i.e., a map of prediction errors for contrast and motion), and the ventral system (i.e., a map of recognition errors in the fovea). Note that values from all three systems are in the range from 0 to 1; input from the superior colliculus is binary, while the other two input maps are continuous. Similarly, input from the ventral system is limited to the fovea, while input from the other two systems spans the entire display.

It is important to note that this processing strategy is consistent with a bottom-up approach, which assumes that shifts in attention are stimulus-driven rather than expectation-driven (Schlesinger, 2003). We return to this issue in the **Discussion**, below.

Training and Testing

Analogous to infants' experience in the car study, the model was first trained for 300 trials on the familiarization event (see Figure 1A). On each trial, the 210 image frames were presented to the model in sequence. Gaze shifts in the model were achieved by determining the most active location in the saliencemap during the current input frame, and shifting the fixation point (i.e., the fovea) to that location prior to the next input frame.

During training, the backpropogation-of-error algorithm was used to modify connection weights in both the dorsal and ventral systems. The mean errors per pixel in the dorsal and ventral systems at the start of training were 0.49 and 0.35, respectively; these fell to 0.01 and 0.09 by the end of training. After 300 training trials, the model was

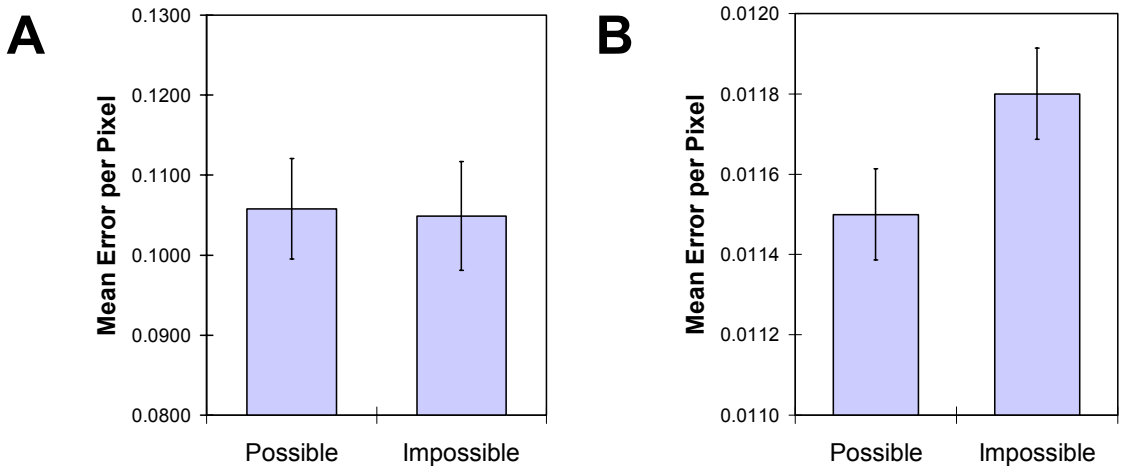

Figure 3: Mean error per pixel during the possible and impossible events, in (A) the ventral and (B) dorsal processing systems (error bars are ±1 SD).

then tested by presenting the possible and impossible events to the network, with all connection weights held constant (i.e., learning was turned off).

Results

We evaluated our two predictions by initializing, training, and then testing the model 20 times. As an analog to looking time in infants, we measured prediction and recognition errors in the dorsal and ventral systems, respectively, during the possible and impossible events.

Figure 3A presents mean error per pixel in the ventral system, during the possible and impossible events. Mean recognition errors in the ventral system during the possible and impossible events were 0.11 and 0.10, respectively. As predicted, there was no significant difference in recognition errors produced by the ventral system to the two test events ($t(38) = 0.44$, p = ns).

Meanwhile, Figure 3A presents mean error per pixel in the dorsal system, during the possible and impossible events. Mean prediction errors were 0.0115 and 0.0118, respectively, during the possible and impossible events. Also as predicted, mean prediction error was significantly higher during the impossible event ($t(38) = 7.26$, $p < .0001$).

As a supplemental analysis, we compared mean error in the dorsal and ventral systems during the familiarization event, at the start and then again at the end of training. Interestingly, at the start of training we found that dorsal error was significantly higher than ventral error ($t(38) = 5.90$, $p < .0001$). However, dorsal prediction error fell more quickly during training than ventral recognition error, and by the end of training dorsal error was significantly lower than ventral error ($t(38) = 40.47$, $p < .0001$). This pattern appears to mirror the developmental trajectory of the ventral and dorsal pathways in humans, with the dorsal pathway developing more rapidly than the ventral pathway (i.e., what before where). As Figure 2B suggests, one possibility for this developmental difference may be that the dorsal path receives relatively sparse inputs (e.g., edges and motion), while the high-resolution inputs of the ventral system appear to carry more detailed information.

Discussion

Our two primary goals in the current paper were to: (1) use the dorsal and ventral pathways as the basis for proposing a decomposition of infants' object representations, and (b) to evaluate this proposal by simulating infants' gaze patterns with the eye-movement model. We described the car study, and reasoned that the location of the box during the possible and impossible events was a critical feature. Therefore, we predicted that the dorsal system (i.e., the where system) would respond to the impossible event as more novel or unexpected (i.e., with higher prediction errors). A second, related prediction was that the ventral system (i.e., the what system) would not be sensitive to where the box was located during the test trials. Analysis of the errors in the ventral and dorsal systems after training and testing the model provided support for both of our predictions.

These results raise an obvious question: Why does placing the box on the track lead to higher (as opposed to lower) prediction errors in the dorsal system? A tentative answer to this question is suggested by the performance of earlier versions of the eye-movement model (e.g., Schlesinger, 2003). In particular, our prior results suggest that over time the trajectory of the car becomes a "special" or privileged region in the display, and novel objects which appear along this trajectory may therefore be more salient. We are currently pursuing this explanation as we continue to test and evaluate the model.

Finally, we note that the current implementation of the eye-movement model focuses on the role of bottom-up or stimulus-driven processing. Our working hypothesis is that bottom-up processing (i.e., prediction errors in the dorsal system) may provide a "preattentive" cue—a sort of subconscious "Hey, what just happened?"—that triggers a more deliberate or top-down analysis of the scene (e.g., Baillargeon, 1995).

References

Baillargeon, R. (1986). Representing the existence and the location of hidden objects: Object permanence in 6- and 8-month-old infants. *Cognition*, *23*, 21-41.

Baillargeon, R. (1993). The object concept revisited: New directions in the investigation of infants' physical knowledge. In C.E. Granrud (Ed.), *Visual perception and cognition in infancy*, pp. 265-315. Hillsdale, NJ: Lawrence Erlbaum.

Baillargeon, R. (1995). A model of physical reasoning in infancy. In C. Rovee-Collier and L.P. Lipsitt (Eds.), *Advances in Infancy Research*, pp. 305-371. Norwood, NJ: Ablex.

Baillargeon, R. (1999). Young infants' expectations about hidden objects: A reply to three challenges. *Developmental Science*, *2*, 115-132.

Baillargeon, R., and DeVos, J. (1991). Object permanence in young infants: Further evidence. Child Development, 62, 1227-1246.

Haith, M.M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behavior & Development*, *21*, 167-179.

Itti, L., & Koch, C. (2000). A saliency-based mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489-1506.

Jacobs, R.A., Jordan, M.I., & Barto, A.G. (1991). Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science*, *15*, 219-250.

Leslie, A.M., Xu, F., Tremoulet, P.D., & Scholl, B.J. (1998). Indexing and the object concept: Developing "what" and "where" systems. *Trends in Cognitive Sciences*, *2*, 10-18.

Mareschal, D., & Johnson, M.H. (2003). The "what" and "where" of object representations in infancy. *Cognition*, *88*, 259-276.

Mareschal, D., Plunkett, K., and Harris, P. (1999). A computational and neuropsychological account of object-oriented behaviours in infancy. *Developmental Science*, *2*, 306-317.

Milner, A.D., & Goodale, M.A. (1995). *The visual brain in action*. New York: Oxford University Press.

Mishkin, M., Ungerleider, L.G., & Macko, K.A. (1983). Object vision and spatial vision: Two central pathways. *Trends in Neuroscience*, *6*, 414-417.

Rueckl, J.G., Cave, K.R., & Kosslyn, S.M. (1989). Why are "what" and "where" processed by separate cortical visual systems? A computational investigation. *Journal of Cognitive Neuroscience*, *1*, 171-186.

Schlesinger, M. (2003). A lesson from robotics: Modeling infants as autonomous agents. *Adaptive Behavior*, *11*, 97-107.

Schlesinger, M., and Barto, A. (1999). Optimal control methods for simulating the perception of causality in young infants. In M. Hahn and S.C. Stoness (Eds.), *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society*, pp. 625-630. New Jersey: Erlbaum.

Schlesinger, M, & Casey, P. (2003). Where infants look when impossible things happen: Simulating and testing a gaze-direction model. *Connection Science*, *15*, 271-280.

Schlesinger, M, and Parisi, D. (2001). The agent-based approach: A new direction for computational models of development. *Developmental Review*, *21*, 121-146.

Schlesinger, M., & Young, M.E. (2003). Examining the role of prediction in infants' physical knowledge. In R. Alterman and D. Kirsh (Eds.), *Proceedings of the Twenty-Fifth Annual Meeting of the Cognitive Science Society* (pp. 1047-1052). Boston: Cognitive Science Society.

Spelke, E.S. (1998). Nativism, empiricism, and the origins of knowledge. *Infant Behavior and Development*, *21*, 181-200.

Steward, O. (2000). *Functional neuroscience*. New York: Springer.