Connection Science
Vol. 18, No. 2, June 2006, 207–216

Taylor & Francis
Taylor & Francis Group

Decomposing infants' object representations: A dual-route processing account

MATTHEW SCHLESINGER*

Brain and Cognitive Sciences Program, Department of Psychology,
Southern Illinois University, Carbondale, IL 62901, USA

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 functional decomposition is supported by the distinction between dorsal and ventral extrastriate visual processing in the primate visual system. This approach is illustrated by presenting 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 (R. Baillargeon, "Representing the existence and the location of hidden objects: object permanence in 6- and 8-month-old infants", *Cognition*, 23, pp. 21–41, 1986.). As expected, the model highlights how the dorsal system is sensitive to the location of a key feature in these events (*i.e.* the location of an obstacle), whereas the ventral system responds equivalently to the possible and impossible events. These results are used to help explain infants' reactions in looking-time studies.

Keywords: Object representations; Dorsal–ventral model; Infant perception

1. 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, three-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 (Baillargeon 1999, Spelke 1998, Haith 1998, Smith 1999).

Much of the research on infants' object knowledge (Baillargeon 1995) has focussed 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 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); (3) to design a computational model that successfully incorporates and exploits these representations.

*Tel.: +1 618-453-3524; Fax: +1 618-453-3563; Email: matthews@siu.edu

Connection Science
ISSN 0954-0091 print/ISSN 1360-0494 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/09540090600768740

In this article, a research strategy for addressing these challenges is presented that is inspired by both structural and functional properties of the primate visual system. Specifically, the extrastriate dorsal and ventral streams are 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 article is organized as follows. In section 2, the properties of the dorsal and ventral pathways are reviewed. The 'car study' (Baillargeon 1986, Schlesinger and Casey 2003), which is used as a platform for developing and testing the model, is then briefly described. Next, key features of the eye-movement model are presented, including how the model is used to simulate infants' gaze patterns in the car study. This article concludes by describing the performance of the model and discussing some of the future directions of this work.

2. 'What-and-where' as task decomposition

The key reason for focussing on the dorsal–ventral distinction as a modeling strategy is that it provides a biologically inspired 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 (Milner and Goodale 1995). Two key features of dorsal processing are a high sensitivity to contrast and to motion (particularly in the periphery) (Steward 2000).

Secondly, 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 (Mishkin *et al.* 1983). Consequently, what processing relies on high-resolution information (particularly from the fovea) (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 are more efficiently accomplished by decomposing the problem into two parallel tasks, identifying objects versus localizing objects, rather than solving both tasks with only one system (Jacobs *et al.* 1991, Rueckl *et al.* 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 where pathway develops first, whereas the what pathway appears to develop more slowly over the first year (Leslie *et al.* 1998, Mareschal *et al.* 1999). Recent evidence also suggests that the what and where streams may not start to become coordinated and integrated until the end of the first year (Mareschal and Johnson 2003).

The what–where distinction also offers an important insight for the study of infants' early object representations. As section 3 highlights, 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 what–where model not only suggests an explicit mechanism for representing salient objects and their locations but is also based on known principles from developmental visual neuroscience.

3. The 'car study'

The car study was designed and first investigated by Baillargeon (1986) and Baillargeon and DeVos (1991). In this study, infants watch a simple mechanical display, in which a car rolls

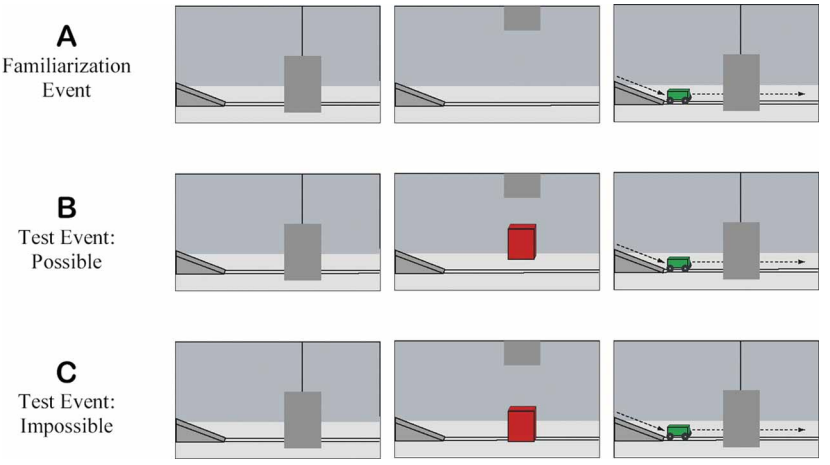

Figure 1. Schematic display of the familiarization (A), possible (B), and impossible (C) events in the car study.

down a ramp, behind a screen and out the other side. Figure 1(A) presents a schematic display of this familiarization event. Note that at the start of the familiarization event, the screen is first 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 (figure 1(B) and (C)). 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 6 months of age, 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. Secondly, 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 what–where 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 are maintained along the what and where paths via persistent neural activity patterns, which provide a memory trace during breaks in contact with an object (*e.g.* occlusion). Secondly, 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.

This argument was pursued by reasoning that because the location of the box is a critical feature in the car study, it should be the where 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, two key predictions were proposed. First, it was predicted that as 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 what system to the possible and impossible events (*i.e.* both events should be equally 'novel'). Secondly, and in contrast, a significant difference was expected in the response of the where system during the test events. Specifically, it was predicted that the where system would show a stronger 'novelty' reaction to the impossible test event.

4. Modeling infants' gaze patterns

The present model is based on a platform that was developed to simulate infants' gaze patterns in looking-time studies (Schlesinger 2003, Schlesinger and Barto 1999, Schlesinger and Parisi 2001, Schlesinger and 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; (3) the production of simulated eye movements that enable the fixation point to move over time.

There are two major innovations in the present model. First, digitized video (rather than computer animation) was used as input to the model, which was recorded in the lab from the same location where infants sit as they view the car study (Schlesinger and Casey 2003). Secondly, the architecture of the model was elaborated to include three processing streams or pathways (*i.e.* the where, what, and also the superior colliculus pathways). Although this approach is similar to a recent model that also simulates the what and where pathways in infants (Mareschal *et al.* 1999), three unique features are: (1) the use of digital video input; (2) production of eye movements; (3) two levels of visual resolution.

4.1 *Model architecture*

Figure 2(A) 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 (sections 4.1.3 and 4.1.4). Figure 2(B) illustrates activity along two of the pathways (*i.e.* what and where) during a sample input frame, after 300 training trials (see Training and testing).

4.1.1 Input. Three events from the car study, *i.e.* familiarization, possible, and impossible (figure 1(A–C)) were produced in the 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 and Casey 2003). The duration of each event was 7 s and each frame was 240 by 180 pixels (in grayscale). The video streams for each event 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 pre-processed with motion and edge filters.

4.1.2 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). This path was included 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 2(A) 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 one.

4.1.3 Where path. To capture the key roles of contrast and motion processing in the where pathway, low-resolution edge and motion frames were combined into a single image and used as input into the where system. Because the where path plays a critical role in spatially oriented actions, a prediction-learning algorithm (*i.e.* forward model) was implemented as a proxy for action guidance. In particular, the task of the where system is to learn to predict the next image frame for each input frame that it receives (for a related approach, see Schlesinger and Young 2003). In particular, a three-layer, fully connected network (1728, 172, and 1728 units, respectively) was employed.

The left side of figure 2(B) illustrates processing along the where pathway during a sample input frame. In particular, whereas the car is visible in the input frame, the where system fails to correctly reproduce it in the output. In this case, note that where 'error' (the absolute value of where output minus where input) is maximal in the region of the car.

4.1.4 What path. Because the what path relies on high-resolution visual input, a 30-by-30 pixel region was sampled from each high-resolution frame, centered on the current fixation point during that frame, as input to the what system. Specifically, the what system was

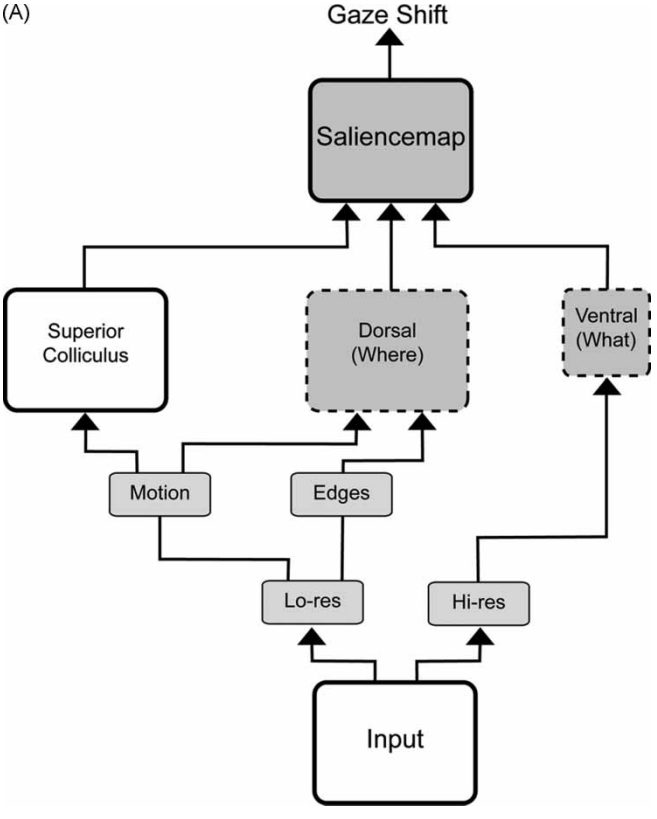

Figure 2. Schematic diagram of the eye-movement model, including (A) model architecture and (B) processing of a sample input frame through the what and where systems after 300 training trials (superior colliculus path not shown).

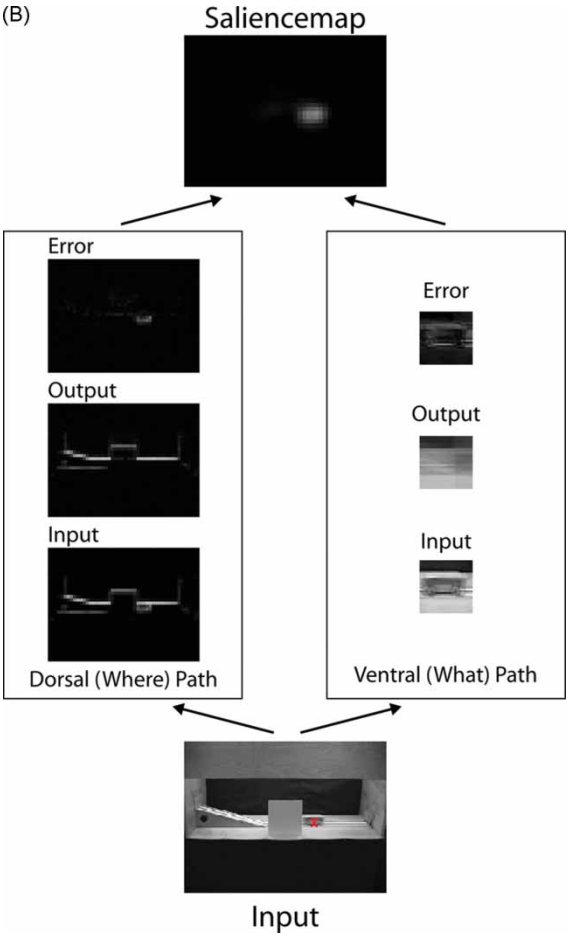

Figure 2. Continued.

implemented as a three-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 what system.

The right side of figure 2(B) illustrates processing along the what pathway during a sample input frame; note that during this frame, the fixation point is centered near the car (*i.e.* just to the right of the screen). As in the where system, the what system also has some difficulty in processing the car after 300 training trials. In particular, while the what 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.

4.1.5 Saliencemap. The last stage of processing in the model is the saliencemap, which pools activations from the earlier pathways into a single, coherent representation (for a related approach, see Itti and Koch 2000). Specifically, the saliencemap sums input from the superior colliculus system (*i.e.* a binary motion map), the where system (*i.e.* a map of prediction errors for contrast and motion), and the what 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, whereas the other two input maps vary continuously. Similarly, input from the what system is limited to the fovea, whereas input from the other two systems spans the entire display.

It is important to note that in designing the model, several features were explicitly chosen to emphasize a bottom-up processing strategy. First, although the what and where systems are both implemented with the use of supervised-learning algorithms, it should be stressed that the target outputs (*i.e.* the training signals) are derived from the model's own sensory experience (*i.e.* it relies on a mismatch between expected and observed sensory inputs). This approach contrasts with the conventional use of supervised learning, in which the training signal is computed on the basis of information that is not directly available to the model or learning agent (for a related discussion, see McClelland 1995, Parisi *et al.* 1990). Secondly, note that where the model fixates is determined by the activity of the saliencemap, which receives its input in a feedforward, bottom-up fashion. In other words, the current approach assumes that shifts in attention are stimulus driven rather than expectation driven (Schlesinger 2003). This issue is returned to in the Discussion.

4.2 *Training and testing*

Analogous to infants' experience in the car study, the model was first trained for 300 trials on the familiarization event (figure 1(A)). On each trial, the 210 image frames were sequentially presented to the model. Gaze shifts in the model were achieved by (1) determining the most active location in the saliencemap during the current input frame, and then (2) 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 what and where systems. The mean errors per pixel in the what and where 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 then tested by presenting the possible and impossible events to the network, with all connection weights held constant (*i.e.* learning was turned off).

5. Results

The two predictions were evaluated by initializing, training, and then testing the model 20 times. As an analog to looking time in infants, prediction and recognition errors in the what and where systems were measured, respectively, during the possible and impossible events.

Figure 3(A) presents mean error per pixel in the what system, during the possible and impossible events. Mean recognition errors in the what 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 what system to the two test events ($t(38) = 0.44$, $p =$ ns).

Meanwhile, figure 3(B) presents mean error per pixel in the where 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 ($t(38) = 7.26$, $p < 0.0001$).

As a supplemental analysis, mean errors in the what and where systems were compared at the start of training and then again at the end of training (recall that errors were normalized as a function of the number of pixels in each system). Interestingly, at the start of training, where error was significantly higher than what error ($t(38) = 5.90$, $p < 0.0001$). However, where

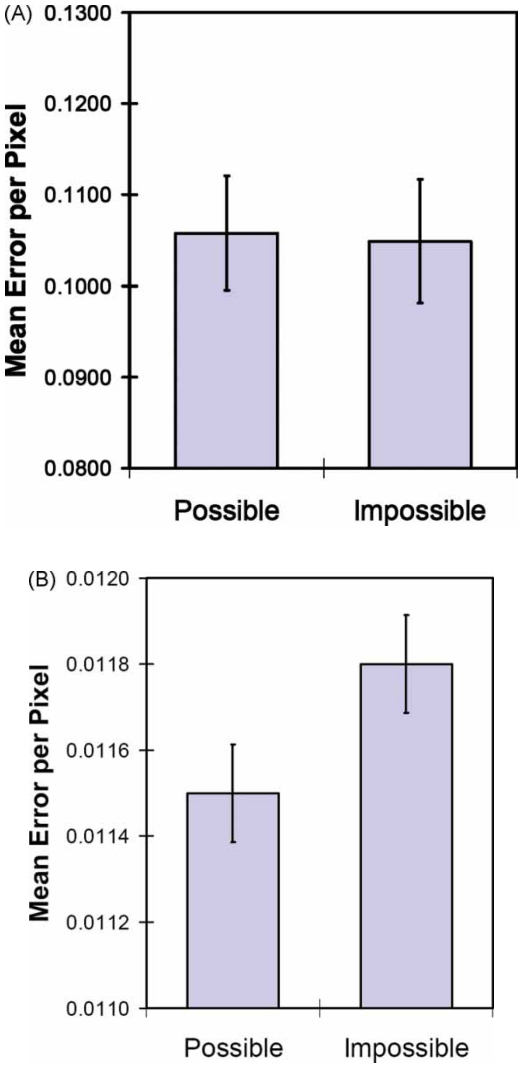

Figure 3. Mean error per pixel during the possible and impossible events, in (A) the what and (B) where processing systems (error bars are ±1 SD).

prediction error fell more quickly during training than what recognition error, and by the end of training where error was significantly lower than what error ($t(38) = 40.47$, $p < 0.0001$). This pattern appears to mirror the developmental trajectory of the what and where pathways in humans, with the where pathway developing more rapidly than the what pathway. As figure 2(B) suggests, one possibility for this developmental difference may be that the where path receives relatively sparse inputs (*e.g.* edges and motion), whereas the high-resolution inputs of the what system appear to carry more detailed information.

6. Discussion

The two primary goals in the current article were (1) to use the what and where pathways as the basis for proposing a decomposition of infants' object representations and (2) to evaluate this

proposal by simulating infants' gaze patterns with the eye-movement model. The car study was described and used to reason that the location of the box during the possible and impossible events was a critical feature. Therefore, it was predicted that the where system should respond to the impossible event as more novel or unexpected (*i.e.* with higher prediction errors) than the possible event. A second, related prediction was that the what system would not be sensitive to differences in the locations of the box during the test events. Analysis of the errors in the what and where systems after training and testing the model provided support for both of these predictions.

These results raise an obvious question: why does placing the box on the track lead to higher prediction errors in the where system? A tentative answer to this question is suggested by the performance of earlier versions of the eye-movement model (Schlesinger 2003). In particular, the prior results suggest that during training, 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. This explanation is currently being pursued as the model continues to be tested and evaluated.

Finally, it is noted that the current implementation of the eye-movement model focusses on the role of bottom-up or stimulus-driven processing. The working hypothesis is that bottom-up processing (*i.e.* prediction errors in the where system) may provide a 'preattentive' cue – a sort of subconscious 'Hey, what was that?' – that triggers a more deliberate or top–down analysis of the scene (Baillargeon 1995).

References

R. Baillargeon, "Representing the existence and the location of hidden objects: object permanence in 6- and 8-month-old infants", *Cognition*, 23, pp. 21–41, 1986.

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

R. Baillargeon, "Young infants' expectations about hidden objects: a reply to three challenges", *Dev. Sci.*, 2, pp. 115–132, 1999.

R. Baillargeon and J. DeVos, "Object permanence in young infants: further evidence", *Child Dev.*, 62, pp. 1227–1246, 1991.

M.M. Haith, "Who put the cog in infant cognition? Is rich interpretation too costly?" *Infant Behav. Dev.*, 21, pp. 167–179, 1998.

L. Itti and C. Koch, "A saliency-based mechanism for overt and covert shifts of visual attention", *Vision Res.*, 40, pp. 1489–1506, 2000.

R.A. Jacobs, M.I. Jordan and A.G. Barto, "Task decomposition through competition in a modular connectionist architecture: the what and where vision tasks", *Cog. Sci.*, 15, pp. 219–250, 1991.

A.M. Leslie, F. Xu, P.D. Tremoulet and B.J. Scholl, "Indexing and the object concept: developing "what" and "where" systems". *Trends Cogn. Sci.*, 2, pp. 10–18, 1998.

D. Mareschal and M.H. Johnson, "The "what" and "where" of object representations in infancy", *Cognition*, 88, pp. 259–276, 2003.

D. Mareschal, K. Plunkett and P. Harris, "A computational and neuropsychological account of object-oriented behaviours in infancy", *Dev. Sci.*, 2, pp. 306–317, 1999.

J.L. McClelland, "A connectionist perspective on knowledge and development", in *Developing Cognitive Competence: New Approaches to Process Modeling*, T.J. Simon and G.S. Halford, Eds, Hillsdale, NJ: Lawrence Erlbaum, 1995, pp. 157–204.

A.D. Milner and M.A. Goodale, *The Visual Brain in Action*, New York: Oxford University Press, 1995.

M. Mishkin, L.G. Ungerleider and K.A. Macko, "Object vision and spatial vision: two central pathways", *Trends Neurosci.*, 6, pp. 414–417, 1983.

D. Parisi, F. Cecconi and S. Nolfi, "Econets: neural networks that learn in an environment", *Network*, 1, pp. 149–168, 1990.

J.G. Rueckl, K.R. Cave and S.M. Kosslyn, "Why are "what" and "where" processed by separate cortical visual systems? A computational investigation", *J. Cogn. Neurosci.*, 1, pp. 171–186, 1989.

M. Schlesinger, "A lesson from robotics: modeling infants as autonomous agents", *Adapt. Behav.*, 11, pp. 97–107, 2003.

M. Schlesinger and A.G. Barto, "Optimal control methods for simulating the perception of causality in young infants", in *Proceedings of the 21st Annual Conference of the Cognitive Science Society*, M. Hahn and S.C. Stoness, Eds, New Jersey: Erlbaum, 1999, pp. 625–630.

M. Schlesinger and P. Casey, "Where infants look when impossible things happen: simulating and testing a gaze-direction model", *Connect. Sci.*, 15, pp. 271–280, 2003.

M. Schlesinger and D. Parisi, "The agent-based approach: a new direction for computational models of development", *Dev. Rev.*, 21, pp. 121–146, 2001.

M. Schlesinger and M.E. Young, "Examining the role of prediction in infants' physical knowledge", in *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, R. Alterman and D. Kirsh, Eds, Boston: Cognitive Science Society, 2003, pp. 1047–1052.

L.B. Smith, "Do infants possess innate knowledge structures? The con side", *Dev. Sci.*, 2, pp. 133–144, 1999.

E.S. Spelke, "Nativism, empiricism, and the origins of knowledge", *Infant Behav. Dev.*, 21, pp. 181–200, 1998.

O. Steward, *Functional Neuroscience*, New York: Springer, 2000.