Developmental Mechanisms in the Perception of Object Unity

Denis Mareschal (d.mareschal@bbk.ac.uk)
Centre for Brain and Cognitive Development; Department of Psychology
Birkbeck College; Malet St
London, WC1E 7HX UK

Scott P. Johnson (spj@psyc.tamu.edu)
Department of Psychology; Texas A&M University
College Station, TX 77843-4235 USA

Abstract

Neonates seem to perceive two ends of a partly occluded rod as two separate objects. However, by 4 months of age infants often appear to perceive a similar stimulus as comprised of a single unified object. Little is known about the mechanisms of development underlying this change. We constructed four connectionist models of how perception of object unity might develop in human infants, based on experience with a variety of visual cues known to be important to infants' performance. After exposure to a simulated visual environment, all the models were able to perceive a partly occluded object as unified. A rich perceptual environment and the presence of units for internal representations were found to improve generalization of acquired unity knowledge. These results lend plausibility to mechanistic accounts of human perceptual development, based on learning the statistical regularities inherent in the normal visual environment.

Introduction

Research exploring the development of object perception often employs simple displays while recording young infants' responses to object properties. For example, the display depicted in Figure 1 appears to adults to consist of a center-occluded rod, moving back and forth behind a nearer box. By 4 months of age, infants appear to perceive such a partly occluded rod as consisting of a single unified object. Earlier studies of the cues that support the perception of object unity concluded that common motion of the rod parts was the primary visual cue used by infants in determining that the rod parts belonged to a common object (Kellman & Spelke, 1983; Kellman, Spelke, & Short, 1986).

However, more recent studies have called this finding into question by systematically varying the cues available in occlusion displays. Three-dimensional depth cues were found to be not necessary for the perception of unity, given that 4-month-olds perceived object unity in a two-dimensional (computer generated) rod-and-box display, in which two rod parts moved above and below a stationary box, against a textured background (Johnson & Náñez, 1995). However, in the absence of a textured background, responses of 4-month-olds appeared to reflect ambiguity with respect to object unity (Johnson & Aslin, 1996). The relatability of the two rod segments (the fact that, if extended, they would meet behind the screen) was also found to be important to infants' perception of unity (Johnson & Aslin, 1996).

Currently, there are few accounts of how this fundamental skill develops. Spelke (1990; Spelke & Van de Walle, 1993) has suggested that young infants' object perception is tantamount to reasoning, in accordance with a set of core principles. However, infants' performance on object unity tasks appears to be strongly dependent on the presence or absence of motion, edge alignment, accretion and deletion of background texture, and other cues, implying that low-level perceptual variables influence the development of veridical object perception, rather than reasoning from core principles (Johnson & Aslin, 1996; Kellman & Spelke, 1983). Evidence from younger infants also casts doubt on accounts based on innate reasoning: Neonates appear to perceive the rod stimulus depicted in Figure 1 as arising from two disjoint objects (Slater et al., 1990). Two-month-olds have been found to perceive object unity, but only with additional perceptual support, relative to displays used with older infants (Johnson, 1997; Johnson & Aslin, 1995).

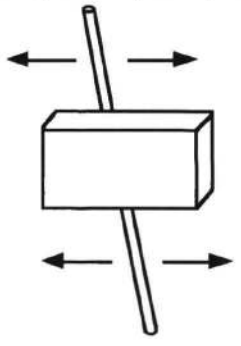


Figure 1. Typical occlusion stimulus

In this paper we explore whether the perception of object unity can be *learned* by experience with objects and events in early infancy. With this goal on mind, we developed a connectionist model that learned to identify a unified partly occluded stimulus from lower-level perceptual cues. The key idea is that when direct perception is not available, a percept of unity can be mediated by an appropriate combination of other supporting cues.

Connectionist models are ideal for modeling learning and development because they develop their own internal representations in response to environmental pressures (Elman et al., 1996). However, they are not simply *tabula rasa* learning machines. The learning that occurs can be strongly determined by innate constraints in the form of specific learning mechanisms or pre-wired connections.

The rest of this paper unfolds as follows. First, we will describe the general model architecture, the input to the model, and what drives learning. Several variations on architecture and training set were tested, but in this paper we report only on the best performing combination. Finally, implications for the development of perception of object unity in human infants will be discussed.

The Basic Architecture

Figure 2 illustrates the basic model architecture. The model receives input via a simple retina. The retinal information is processed by separate encapsulated modules. Each module identifies the presence of one of the following cues: (a) motion, (b) texture deletion and accretion, (c) t-junctions, (d) co-motion (i.e., simultaneous motion) in the upper and lower halves of the retina, (e) common motion in the upper and lower halves of the retina, (f) co-linearity of objects in the upper and lower halves of the retina, (g) the relatability of objects in the upper and lower halves of the retina (i.e., whether the objects' edges would meet if extended behind the occluder).

Unity is also a primitive, like the other cues, in that the network can immediately perceive it (via direct perception). Indeed, when testing the perception of unity in human infants, researchers assume that infants can distinguish single rods from disjoint rod parts. In the absence of direct perception (i.e., when the object(s) are partly occluded) the perception of unity is mediated by its association with other (directly perceivable) cues.

We do not wish to make the claim that a mediated route is unique to the percept of unity. In the brain, there is likely a highly complex and interactive network of connections allowing any number of not-directly-perceivable cues to be indirectly computed from the activation of other directly-perceivable cues. However, in the interest of clarity, we have considered only the one mediated route.

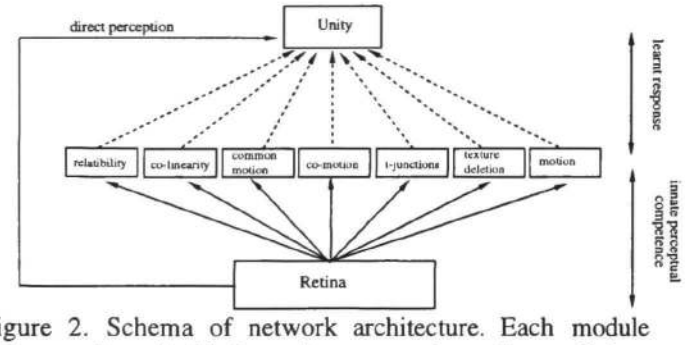
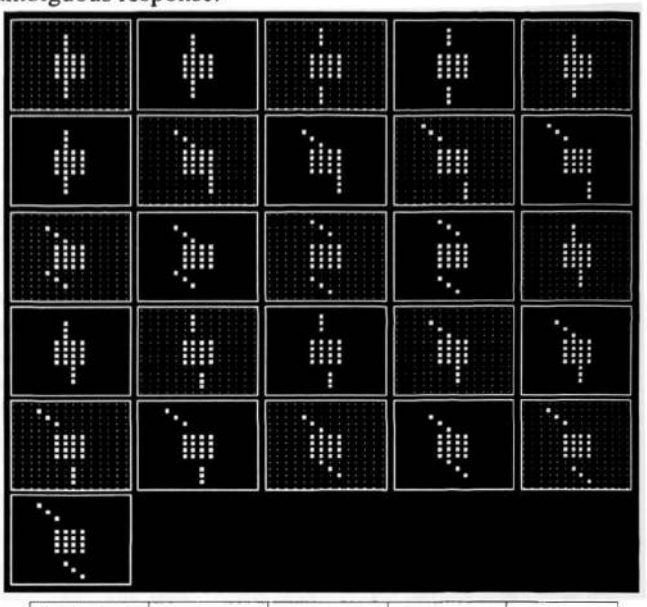


Figure 2. Schema of network architecture. Each module processes the retinal information separately and in parallel.

The bottom half of the network embodies innate abilities. We assume that neonates are able to perceive the components of each of these cues. Indirect evidence suggests that this is the case (Slater, 1995). There is no learning in any of these encapsulated modules. The top half of the network embodies the learning that can occur through interactions with the environment. The models discussed below illustrate several ways in which architectural and environmental constraints can be combined to guide the learning of object unity.

Input to the Model

The network "sees" a series of images from the world and responds with whether a perceived object is unified or not. The response is coded across two output units: (+1, -1) signifies that the object is unified; (-1, +1) signifies that the object is NOT unified. (+1, +1) or (-1, -1) are interpreted as an ambiguous response.



1: motion, texture, t-junction, co-motion, common motion, co-linearity, relatability	2: motion, t-junction, co-motion, common motion, co-linearity, relatability	3: motion, texture, co-motion, common motion, co-linearity, relatability	4: motion, co-motion, common motion, co-linearity, relatability	5: texture, t-junction, co-linearity, relatability
6: t-junction, co-linearity, relatability	7: motion, texture, t-junction, co-motion, common motion, relatability	8: motion, t-junction, co-motion, common motion, relatability	9: motion, texture, co-motion, common motion, relatability	10: motion, co-motion, common motion, relatability
11: motion, texture, t-junction, co-motion, common motion, co-linearity	12: motion, t-junction, co-motion, common motion, co-linearity	13: motion, texture, common motion, co-linearity	14: motion, co-motion, common motion, co-linearity	15: motion, texture, t-junction, co-linearity
16: motion, texture, t-junction, co-linearity	17: motion, texture, co-linearity	18: motion, co-linearity	19: motion, texture, t-junction	20: motion, t-junction
21: motion, texture	22: motion	23: motion, texture, t-junction, co-linearity	24: motion, t-junction, co-linearity	25: motion, texture, co-linearity
26: motion, co-linearity				

Figure 3. Complete set of 26 possible occlusion events. The cues present in the display are listed in the corresponding position of the table.

The input retina consists of a 196-bit vector mapping all the cells on a 14x14-unit grid. In the middle of the grid is a 4x4-unit occluder. All units corresponding to the position of the screen are given a value of 1. When background texture is required, all other units on the retina are given a value of 0.0 or 0.2, depending on the texture pattern. Units with values of 0.2 correspond to position on which there is a texture element (e.g., a dot). Units corresponding to the position of an object (i.e., rod or occluder) are given a value of 1.0. Figure 3 shows a snapshot taken from the "ambiguous" portion of all 26 events in the environment.

An event is made up of a series of snapshots like this one in which the rod moves progressively across the retina. All events begin with the object moving onto the retina from the side. We will call this the *unambiguous* portion of the event. The object then moves across the retina, passing behind the area occupied by the occluding screen. We will call this the *ambiguous* portion of the event. Finally, the

object reappears on the other side of the screen and continues off the retina.

All events except 5 and 6 involve motion. The presence of texture, t-junctions, relatability and co-linearity are varied systematically. All events with motion involve motion in the upper and lower half of the retina (co-motion) but only half of those involve common motion. This leads to a total of 26 possible events.

The Perceptual Modules

These modules are not intended to model closely human neurophysiology. Although the modules embody general neural computational principles of summation, excitation, inhibition and local computation, they are also tailored to the specific nature of our networks' visual experience. Thus, they are not general models of the human visual system. However, they do embody some of the basic principles believed to underlie the computation of various visual cues (see Spillman & Werner, 1990 for a review). In essence, they are neurally plausible information processing modules.

All the modules compute the presence or absence of a relevant cue from the retinal image. The principles on which the modules function are as follows:

- *Motion detection module*

Takes the current retinal image and compares it to the previous retinal image. If there is a difference between the images, then there has been motion. If not, then there has not been any motion.

- *Texture module*

Counts the number of texture dots in the input image and compares it to the number of dots in the previous image. If there is a difference in the number of dots in the two images, then there has been deletion and/or accretion of texture elements.

- *T-junction module*

Focuses on the area immediately above and below the edge of the occluding screen and computes whether there is a gap everywhere along the edge.

- *Co-motion module*

Splits the retina into two halves and computes whether there is motion both in the upper half and in the lower half.

- *Common-motion module*

Splits the retina into two halves and computes whether there is the same direction of motion in the upper and the lower halves.

- *Co-linearity module*

Computes the tangent of the angle that the object's axis of principle length makes with the horizontal for both the upper and lower halves of the retina and compares these two.

- *Relatability module*

Computes whether the extension of the axis of principle length for objects in the upper and lower halves of the retina will intersect.

Although alignment has been manipulated as a cue in some infant studies, note that two objects are aligned *if and only if* they are co-linear and relatable. Thus, co-linearity and relatability are more primitive cues than alignment in the sense that the later cannot be computed without computing the former (at least implicitly), whereas the converse is not true: Both co-linearity and relatability can be computed independently of alignment.

As an example, the motion detection module is illustrated in Figure 4. This module takes in the retinal input and returns 1 if there is motion on the retina or 0 if there is no motion on the retina. The basic principle of the module is to take the current retinal image and compare it to the previous retinal image. If there is a difference between the images, then there has been motion. If not, then there has not been any motion.

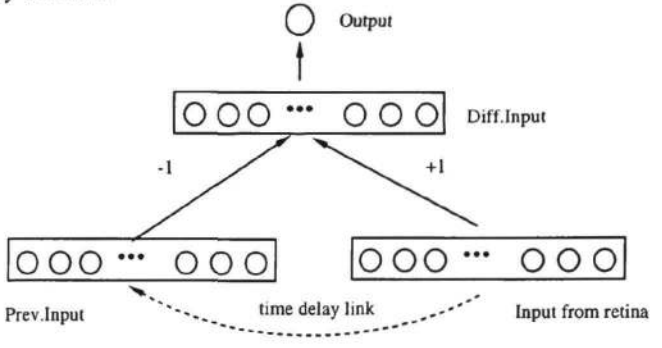


Figure 4. Schema of motion detection module

The Input is copied to a memory buffer (Prev.Input). A layer of hidden units (Diff.Input) computes the unit by unit difference between the current input and the previous input. The output unit then sums the activity across the hidden layer and returns 1 if there are any non-zero values or 0 if there are no non-zero values.

What drives learning?

Learning is partly driven by a feedback signal from the environment and partly driven by memory. When the object is visible, the environment provides immediate feedback about the unity of the object via the direct perception link. When the object is not completely visible, the environment cannot provide feedback about the unity of the object. To overcome this problem, the model has a short-term, rapidly decaying memory that encodes unity information obtained from direct perception (i.e., a kind of visual memory). Immediately following occlusion there is a clear trace of the state of the rod before occlusion. After a short delay, that information disappears and can no longer be used for learning.

This relation between direct perception and memory is embodied in the target signal used for training the weights:

$$T_i(t) = E_i(t) + \mu . T_i(t-1) \qquad (1)$$

with $-1 < T_i < +1$, $0 < \mu < 1$, and $E_i = 0.0$ when the rod is occluded.

E_i is the unity signal obtained from the environment by direct perception for output i, and μ is a parameter controlling the depth of memory. When $E_i = 0.0$ (i.e., there is no direct percept of unity), the target (T_i (t)) is derived entirely from the memory component $\mu.T_i(t-1)$, the second term in the right-hand side of the equation.

An interesting component of the model's performance is the mediated route's ability to predict whether a test event corresponds to a single unified object or to two disjoint objects. Network performance can be assessed either when direct perception is possible, or when it is not possible (i.e., on events 1, 2, 5, 6, 7, and 8 in Figure 3).

When direct perception is not possible, the model's prediction of unity (via the mediated route) can be compared to the modeler's knowledge of whether the test event arises from a unified object or not. When direct perception is possible, the model's prediction of unity (via the mediated route) can be compared to the signal coming from direct perception.

The degree to which the model's prediction is correct when direct perception is NOT possible reflects how well the model is able to respond to incomplete information. This can then be compared to infants' performance when faced with the same ambiguous stimuli. The degree to which the prediction is correct when direct perception IS possible reflects how well the network has extracted general information about objects that applies across its entire learning environment.

Model Performance

Four combinations of architecture and environment were explored. The models either had no hidden units between the output of the modules and the unity response units (see Figure 2), or they had three hidden units between the output of the modules and the unity response units. (The addition of hidden units provides the model with the power to develop internal representations of cue combinations and to represent to non–linearly separable cue relations, such as the exclusive-or, of a set of cues.) In addition, the models were trained either with a *basic* but ecologically plausible set of events (events 1, 2, 3, and 4), or were trained with an *enriched* set of events that sampled more evenly the space of possible object events (events 1, 2, and 17 though 22).

In the interest of brevity we only report the models' performance in the best conditions (complete results will be reported in a future paper). For the moment, it suffices to note that the presence of hidden units and a richly varying environment led to the best generalization performance.

Three hidden units were added between the perception modules and the output response units. The training environment was enriched to capture the fact that there are far more examples of disjoint objects in the real world than unified but occluded objects. Weights were adjusted by applying the backpropagation algorithm with learning rate = 0.5, momentum = 0.03, logistic activation functions, and memory (μ) = 0.4. The results are based on 10 replications with different random initial weights.

The networks very quickly learned (by 10 epochs) to perceive one or two objects during the unambiguous portion of the events. The unambiguous portion of the events corresponds to the time when the rod(s) were moving across the retina and had not yet reached a position of partial occlusion (one rod part above and the other below the occluder). That is, the rod or rod parts were directly visible.

Figure 5 shows the networks' performance when tested with the ambiguous portion of events 1, 3, 5, and 7, events used to test infants (see Johnson, 1997). Note that only event 1 was part of the original training set.

Veridical perception of a single unified object (event 1) was apparently rather difficult to learn (see Figure 5, top left). Over the first 4000 epochs, the networks perceived event 1 as arising from two disjoint objects. Then, from 5000 to 8000 epochs the majority of networks gradually came to perceive the event as arising from a single object. In contrast, the networks very quickly learned (by 100 epochs) to perceive two objects when the event was actually produced by two objects (event 3; see Figure 5, top right).

There was a different developmental profile in the absence of motion (event 5; see Figure 5, bottom left). Up to epoch 500, the networks perceived this event as arising from two disjoint objects. From epoch 1000 onwards, the networks consistently perceived the event as arising from a single unified object. When the object segments were misaligned but relatable (event 7; see Figure 5, bottom right), the pattern of development was rather different. Throughout development, the networks tended to perceive this event as arising from disjoint objects. At different times, only up to 20% of networks perceived a unified object, whereas the rest perceived two disjoint objects.

Network performance on these four events matches human performance very well (see Johnson, 1997 for review). Initially, the single object depicted in Figure 1 (event 1) was perceived as two disjoint objects (similar to human neonates). There was then a transition period in which either of two responses resulted. After more extensive training, the display was perceived as arising from a single object (similar to older infants and adults). Moreover, human neonates, like these networks, have been found to perceive separate rod parts undergoing common motion (event 3) as consisting of separate objects, after little exposure. Finally, the networks tended to perceive objects with misaligned edges as disjoint (event 7), similar to infants and adults.

The networks were effective in generalizing their "knowledge" to the complete set of test events. By the end of training, they responded correctly to 23 of the 26 test events when tested with the ambiguous portion of the event. That is, the mediated route made incorrect predictions on 3 events (events 22, 23 and 24). The response to events 23 and 24 were altogether incorrect, whereas half the networks provided the correct response to event 22 and half provided an incorrect response. Moreover, the networks performed very well on the unambiguous, visible segments of the trajectories. The mediated route produced the correct percept on 24 of the 26 events. In particular, the networks failed to respond appropriately on events 20 and 22. In the former case, four networks correctly predicted two objects while six predicted a single object, and in the latter case six networks correctly predicted two objects while four predicted a single object.

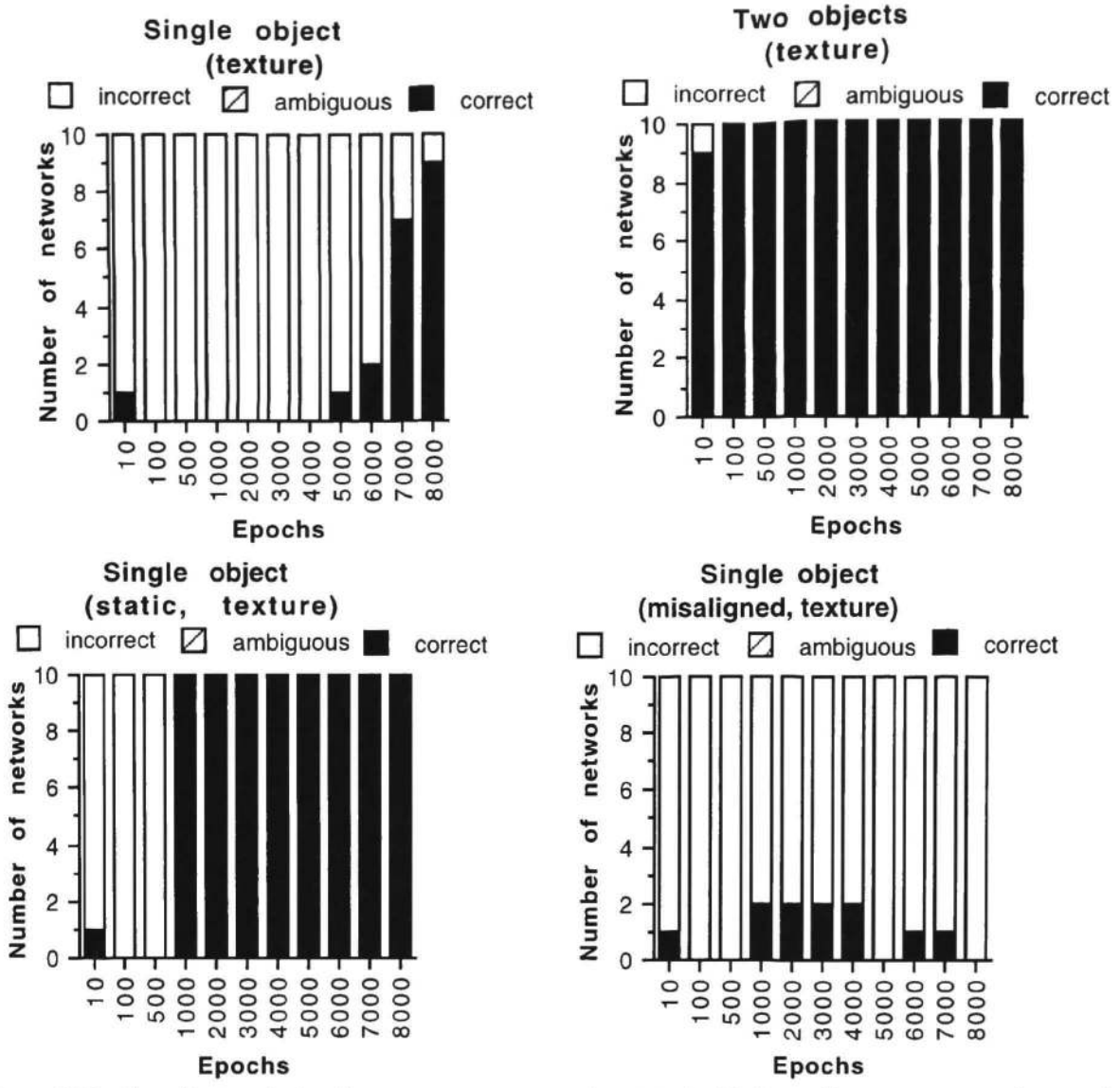


Figure 5. Number of network showing a correct response when tested with the ambiguous segment of events 1, 3, 5, and 7.

An examination of the internal representations developed across the hidden units allows us to explain these responses (Table 1). All three hidden units are used to encode the cue information, but the units have quite different effects on the output responses. Hidden unit 2 is strongly associated with the percept of one object whereas hidden units 1 and 3 are strongly associated with the percept of two objects. Hidden unit 3 has about twice as much impact as hidden unit 1.

The presence of T-junctions is strongly associated with hidden unit 2, and thereby with the percept of unity. Along this dimension, it is the dominant feature. Relatability and common motion are also positively associated with hidden unit 2 (and therefore to the percept of unity), but to a much weaker extent.

Hidden unit 3 is positively associated with all the cues. This means that the unit will almost always be active, whatever the percept. However, because the impact of hidden unit 3 on the output response is less than that of hidden unit 2 (e.g., -2.797 vs. 3.183), the activation of hidden unit 1 will determine which way the network responds. If Hidden unit 1 is active, then the total activation sent to the outputs will produce a "two object" response (activation of the not-unified output node) whereas if the Hidden unit 1 is not active, the total activation sent to the outputs will produce a "one object" response (activation of the unified output node).

Table 1. Connection weights in a representative network

	Relat-ability	Co-linearity	Common-motion	Co-motion	T-junction	Texture deletion	Motion	Bias unit
Hidden unit 1	-0.174	-1.178	-0.659	0.749	-0.237	-0.477	1.513	-0.266
Hidden unit 2	0.431	-0.861	0.555	-1.990	2.358	-0.918	-1.559	3.991
Hidden unit 3	0.825	2.058	1.741	0.208	2.491	0.946	1.010	-2.007

	Hidden unit 1	Hidden unit 2	Hidden unit 3	Bias
unified node	-1.601	3.183	-2.797	-0.101
NOT unified node	1.600	-3.183	2.769	0.101

Hidden unit 1 is positively associated with Motion and Co-motion. Thus if both of these are present, the unit will tend to fire and the response will tend toward the signalling of two objects. If either Motion or Co-motion is absent, hidden unit 1 will be weakened in activity and the network's response will tend toward two objects.

Discussion

Outcomes of these models suggest that perception of object unity can be *learned* rapidly through interaction with the environment. The networks respond to the statistical regularities in the environment: *No prior object representations are required.*

The models can be used to predict the type of percept that will arise from each of the conditions above. Rather than appealing to "core principles" that guide inferences about object unity, the resolution of ambiguous stimuli relies on the previous association of lower-level cues with direct percepts of unity. That is, a strong prediction of the model is that experience viewing unoccluded objects that are progressively occluded and unoccluded lies at the heart of learning to resolve ambiguous stimuli.

In these models, the use of backpropagation is not necessarily crucial; in principle, any algorithm that permits multi-layer learning would suffice (e.g., O'Reilly, 1998). However, there is a need for hidden units (the power for internal re-representation) for proper generalization of knowledge.

Finally, we believe that connectionist models are an effective means of investigating outstanding questions in developmental psychology, in this case by providing a mechanistic account of how learning to perceive object unity could occur.

References

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT Press.

Johnson, S. P. (1997). Young infants' perception of object unity: Implications for development of attentional and cognitive skills. *Current Directions in Psychological Science, 6,* 5-11.

Johnson, S. P., & Aslin, R. N. (1996). Perception of object unity in young infants: The roles of motion, depth, and orientation. *Cognitive Development, 11,* 161-180.

Johnson, S. P., & Aslin, R. N. (1995). Perception of object unity in 2-month-old infants. *Developmental Psychology, 31,* 739-745.

Johnson, S. P. & Náñez, J. E. (1995). Young infants' perception of object unity in two-dimensional displays. *Infant Behavior and Development, 18,* 133-143.

Kellman, P. J., & Spelke, E. S. (1983). Perception of partly occluded objects in infancy. *Cognitive Psychology, 15,* 483-524.

Kellman, P. J., Spelke, E. S., & Short, K. R. (1986). Infant perception of object unity from translatory motion in depth and vertical translation. *Child Development, 57,* 72-86.

O'Reilly, R. C. (1998). Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences, 2,* 455-462.

Slater, A. (1995). Visual perception and memory at birth. In C. Rovee-Collier and L. P. Lipsitt (Eds.), *Advances in infancy research* (Vol. 9). Norwood, NJ: Ablex.

Slater, A. M., Morison, V., Somers, M., Mattock, A., Brown, E., & Taylor, D. (1990). Newborn and older infants' perception of partly occluded objects. *Infant Behavior and Development, 13,* 33-49.

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science, 14,* 29-56.

Spelke, E. S., & Van de Walle, G. (1993). Perceiving and reasoning about objects: Insights from infants. In N. Eilan, R. A. McCarthy, & B. Brewer (Eds.), *Spatial representation: Problems in philosophy and psychology.* Oxford: Blackwell.

Spillman, L. & Werner, J. S. (1990). *Visual perception. The neuropsychological foundations.* London, UK: Academic Press.