A Self-Organizing Neural Network That Learns to Detect and Represent Visual Depth from Occlusion Events

Jonathan A. Marshall and Richard K. Alley
Department of Computer Science, CB 3175, Sitterson Hall
University of North Carolina, Chapel Hill, NC 27599-3175, U.S.A.
marshall@cs.unc.edu, alley@cs.unc.edu

Abstract. Visual occlusion events constitute a major source of depth information. We have developed a neural network model that learns to detect and represent depth relations, after a period of exposure to motion sequences containing occlusion and disocclusion events. The network's learning is governed by a new set of learning and activation rules. The network develops two parallel opponent channels or "chains" of lateral excitatory connections for every resolvable motion trajectory. One channel, the "On" chain or "visible" chain, is activated when a moving stimulus is visible. The other channel, the "Off" chain or "invisible" chain, is activated when a formerly visible stimulus becomes invisible due to occlusion. The On chain carries a predictive *modal* representation of the visible stimulus. The Off chain carries a persistent, *amodal* representation that predicts the motion of the invisible stimulus. The new learning rule uses disinhibitory signals emitted from the On chain to trigger learning in the Off chain. The Off chain neurons learn to interact reciprocally with other neurons that indicate the presence of occluders. The interactions let the network predict the disappearance and reappearance of stimuli moving behind occluders, and they let the unexpected disappearance or appearance of stimuli excite the representation of an inferred occluder at that location. Two results that have emerged from this research suggest how visual systems may learn to represent visual depth information. First, a visual system can learn a nonmetric representation of the depth relations arising from occlusion events. Second, parallel opponent On and Off channels that represent both modal and amodal stimuli can also be learned through the same process.

1 Introduction: Perception of Occlusion Events

During motion, visual objects undergo substantial changes in appearance. They change size, shape, and position with respect to the background (FIGURE 1). They even occasionally disappear behind other objects (FIGURE 1C) and reappear in a new position (FIGURE 1D).

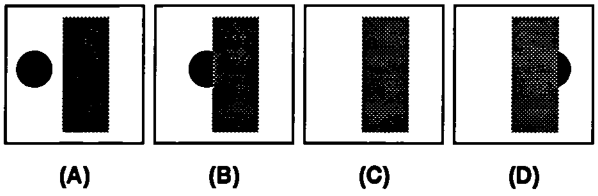

Figure 1: Changes of appearance during an occlusion event. (A) Image frame containing two objects. Circle is moving toward the right. (B) Circle becomes partially occluded, and its image shape changes. (C) Circle becomes completely occluded. (D) Circle begins to emerge on the other side of the occluder, and its shape is again changed.

Our visual systems make highly effective use of these changes to deduce the depth relations among objects. For instance, the occlusion sequence in FIGURES 1B–1D is typically judged to indicate that the rectangle is nearer in depth than the circle. Evidence from psychophysics (Kaplan, 1969; Nakayama & Shimojo, 1992; Nakayama, Shimojo, & Silverman, 1989; Shimojo, Silverman, & Nakayama, 1988, 1989; Yonas, Craton, & Thompson, 1987) and neurophysiology (Frost, 1993, personal communication) suggests that the process of determining relative depth from occlusion events operates at an early stage of visual processing.

(Marshall, 1991) describes evidence that suggests that the same early processing mechanisms maintain a representation of temporarily occluded objects for some amount of time after they have disappeared behind an occluder, and that these representations of invisible objects interact with other object representations, in much the same manner as do representations of visible objects. For example, Shimojo, Silverman, & Nakayama (1988) describe a way in which our visual mechanisms for processing motion information and stereo depth information interact despite the temporary occlusion of a moving object in one or both eyes.

Below we describe how a visual system can *learn* to detect and represent depth relations, after a period of exposure to occlusion and disocclusion events. We use a self-organizing neural network model that exploits the visual changes that occur at occlusion boundaries to form a mechanism for detecting and representing relative depth information. The network's learning is governed by a new set of learning and activation rules.

2 The Predictivity Principle: A Heuristic for Choosing Learning Rules

Our analysis is derived from the following visual *predictivity principle*, which we postulate as a fundamental principle of neural organization in visual systems: *Visual systems represent the world in terms of predictions of its future appearance, and they reorganize themselves to generate better predictions.* If the predictivity principle were satisfied (i.e., a visual system generates perfect predictions of the appearance of its view of the world, down to the last image detail), then clearly we could infer that the visual system possessed an excellent representation or model of the actual state of the visual world. Albus (1991) described a version of a predictivity principle, in which differences between predicted and observed input can lead to updates of an internal world model, so that better predictions can be generated later.

The predictivity principle asserts that visual systems should represent the world in terms of *predictions*. For instance, whenever an object is detected as moving in a certain direction, its motion should be represented in terms of a prediction that after a given delay, the object will appear at a certain location, farther along the motion trajectory.

Marshall (1989, 1990) describes how a neural network model can learn to represent visual motion sequences in terms of prediction signals transmitted along time-delayed lateral excitatory connections. The outputs of any time-delayed connection can be considered a prediction, since these connections transmit information about past scenes to the neurons processing the present sensory inputs.

The predictivity principle further asserts that whenever a prediction fails, a learning rule should be triggered to alter the visual system's world model so that a better prediction would be generated if the same scene were to arise again. For instance, suppose the prediction that the moving object will appear at a certain location fails (e.g., because the object moves behind an occluder). The visual system's motion model should then be altered so that it would have predicted more accurately the disappearance of the object if the same scene were replayed.

2.1 Predictivity Implies Prediction of Occlusion and Disocclusion Events

Taken as a starting point, this simple yet powerful predictivity principle has surprisingly rich implications. Suppose that we want to design a visual system that generates accurate predictions of the occlusion sequence in FIGURES 1A–1D. If the depth relation between the rectangular occluder and the moving circle is known, then it is easy to predict the disappearance of the circle; the system would have to know only the location and velocity of the circle and of the occluder.

The reappearance of the circle (FIGURE 1D) is harder to predict; FIGURE 1C contains no direct evidence that a circle is about to appear anywhere. Nevertheless, if the visual system *can* predict accurately the reappearance of the circle, then it would more completely satisfy the predictivity principle. We therefore conclude that the predictivity principle implies that the visual system *should* predict disocclusion events like the reappearance of the circle in FIGURE 1D, as well as occlusion events.

2.2 Prediction of Disocclusion Implies Representation of Invisible Objects

How can a visual system predict disocclusion events? Somehow it must maintain a representation of the information that it will need. The representation must carry the properties of the occluded object, such as its position, velocity, and shape. Since no information about the occluded object is directly available while it is occluded, this representation must originate before it becomes occluded. This representation must interact with information about the occluder so that when the occluded object reaches the end of the occluder, a prediction that it will become visible again is generated.

If the occluded object is part of a group of objects, some of which are visible, then the motion of the visible objects in the group can control the perceived motion of the invisible object (Anstis & Ramachandran, 1986; Ramachandran, Inada, & Kiama, 1986). Therefore, we argue that the representation of the occluded object can be controlled (e.g., its direction altered) in mid-course, even while the object is invisible. This argument favors a persistent form of representation (the inferred invisible object is represented as a real object, with properties of motion, shape, position, etc.), rather than a ballistic form (with a representation impervious to alteration once it has been launched) (Marshall, 1991).

2.3 Representation of Invisible Objects Implies Separate Representations for Visible and Invisible

A consequence of having a representation of invisible objects is that the representations for visible and invisible objects must be separate or distinguished from each other, so that the visual system can generate separate predictions in both cases. Otherwise, the visual system could not determine whether its representations predict visibility or invisibility, which would contravene the predictivity principle. Thus, simple single-channel prediction schemes like the one described by Marshall (1989, 1990) are inadequate to represent occlusion and disocclusion events.

3 A Grounded System for Learning to Predict Visibility and Invisibility

A visual system that could predict the visibility and invisibility of objects undergoing occlusion and disocclusion would *ipso facto* constitute a representation of the depth relations between the occluded and occluding objects. We considered many mechanisms whereby a visual system could *learn* to generate such predictions.

The basic structure of our networks is one in which a stage of neurons receives a variety of inputs. Some neurons receive strong *bottom-up excitatory* connections from preprocessing stages; these connections transmit information about actual image data. Some neurons receive strong *time-delayed lateral excitatory* connections from other neurons within the same stage; these connections transmit priming excitation from neurons activated at one moment to the neurons predicted to be active at the next moment. All neurons receive *lateral inhibitory* connections from some other neurons; these serve to enforce competitive decision-making in the interpretation of visual input patterns.

Within this scheme, there might exist three classes of neurons at each spatial position in the network. Class I neurons receive both strong bottom-up excitatory and strong lateral excitatory connections (FIGURE 2A). These neurons respond preferentially to events where a visual object moves as predicted, without being occluded or disoccluded.

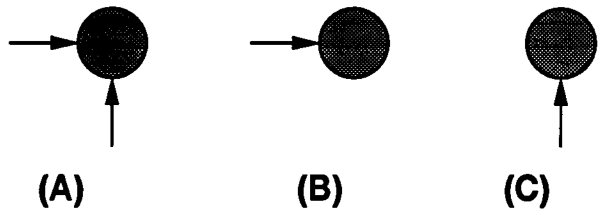

Figure 2: Three classes of neurons for detecting occlusion events. (A) Class I neurons receive both bottom-up and lateral afferents. (B) Class II neurons receive primarily lateral afferents. (C) Class III neurons receive primarily bottom-up afferents.

Class II neurons receive strong lateral excitatory but weak bottom-up excitatory connections (FIGURE 2B). These neurons respond preferentially to events where a moving visual object is predicted (lateral excitation) to appear farther along its trajectory but fails to appear at the predicted place (lack of bottom-up excitation). Such excitation patterns typically arise during occlusion events.

Class III neurons receive strong bottom-up excitatory but weak lateral excitatory connections (FIGURE 2C). These neurons respond preferentially to events where a visual object freshly appears (bottom-up excitation) without having been predicted (lack of lateral excitation). Such excitation patterns typically arise during disocclusion events.

3.1 Learning Separate Representations for Visible and Invisible Implies a Grounded Learning System

Hebbian-type learning rules are attractive because of their simplicity. Although we considered standard variants of Hebbian rules in our quest for a visual system that can learn to represent depth information, we rejected them because they were subject to a severe theoretical problem. If a Class II neuron, which is activated by lateral excitation alone, excites other neurons and causes them to become active without bottom-up excitation, then a Hebbian-type learning rule would cause its lateral excitatory connections to the other neurons to strengthen, and it would cause the bottom-up inputs to the other neurons to weaken. The other neurons would thereby tend to be converted to Class II neurons. Thus, if the network contains Class II neurons, then Hebbian-type rules have no way to prevent *all* neurons within the same network stage from becoming Class II neurons. If that happened, then the network stage would be divorced from actual visual input flowing from prior stages (FIGURE 3B). Neuron activations within the stage would propagate in an uncontrolled, "hallucinatory" manner.

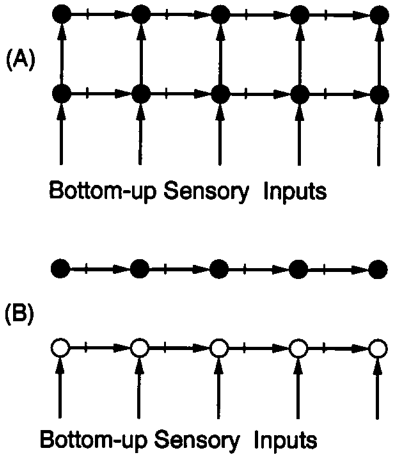

Figure 3: Hallucination problem. (A) The learning rules must be chosen so that bottom-up image information can flow to every layer, e.g., via connections indicated by vertical arrows. (B) Otherwise, some layers might learn strong lateral connections at the expense of bottom-up connections. Such layers would become disconnected from actual visual input signals.

Clearly, that would be a failure. We sought a learning system that would prevent hallucinatory networks from developing. The rules that we chose were thus required to be *grounded* – i.e., to keep some neurons supplied with actual bottom-up input.

3.2 Using Disinhibition to Control Learning of Occlusion Relations

In this paper we will outline one of the methods that we investigated for learning occlusion relations. Other methods may work as well.

Our method involves extending the EXIN (excitatory+inhibitory) learning scheme described by Marshall (1992, 1993). The EXIN scheme uses a variant of a Hebbian rule to govern learning in the bottom-up and time-delayed lateral excitatory connections, plus an anti-Hebbian rule to govern learning in the lateral inhibitory connections.

We extend the EXIN rules by letting inhibitory connections transmit *disinhibitory* signals under certain regulated conditions. The disinhibitory signals produce an excitatory effect at their target neurons.

Our new disinhibition rule specifies that *when a neuron has strong, active lateral excitatory input connections and strong but inactive bottom-up input connections, then it tends to emit disinhibitory signals* *along its output inhibitory connections.* The disinhibitory signals tend to excite the recipient neurons and enable them to learn. The disinhibition rule can be expressed as differential equation governing neuron activation and implemented in a computational simulation.

During continuous motion sequences, without occlusion or disocclusion, the system operates similarly to a system with the standard EXIN learning rules: lateral excitatory "chains" of connections are learned across sequences of neurons along a motion trajectory. A moving visual object activates neurons along a chain; each activated neuron transmits predictive lateral excitatory connections to other neurons farther along the chain.

However, during occlusion events, some predictive lateral excitatory signals reach neurons that have strong but inactive bottom-up excitatory connections. When this occurs, the new disinhibition rule comes into play. The neurons reached by this excitation pattern emit disinhibitory rather than inhibitory signals along their output inhibitory connections. There then exists a neuron that receives a larger amount of disinhibition combined with lateral excitation than other neurons. That neuron becomes more active than other neurons and begins to suppress the activity of the other neurons via lateral inhibition.

In other words, a neuron that represents a visible object causes some *other* neuron to learn to represent the object when the object becomes invisible. Thus, the representations of visible objects are protected from erosion by occlusion events. Moreover, the representations of invisible objects are allowed to develop only to the extent that the neurons representing visible objects explicitly disclaim the "right" to represent the objects. These properties prevent the network from losing contact with actual bottom-up visual input and thereby help it avoid the hallucination problem.

Our system initially contains a homogeneous stage of neurons that receive motion input signals from prior stages. When the system is exposed to many motion sequences containing occlusion and disocclusion events, the stage gradually undergoes a self-organized *bifurcation* into two distinct pools of neurons, as shown in FIGURE 4. These pools consist of two parallel opponent channels or "chains" of lateral excitatory connections for every resolvable motion trajectory. One channel, the "On" chain or "visible" chain, is activated when a moving stimulus is visible. The other channel, the "Off" chain or "invisible" chain, is activated when a formerly visible stimulus becomes invisible, usually due to occlusion. The bifurcation may be analogous to the activity-dependent stratification of cat retinal ganglion cells into separate On and Off layers, found by Bodnarenko and Chalupa (1993).

The On chain (Class I neurons) carries a predictive *modal* representation of the visible stimulus. The Off chain (Class II neurons) carries a persistent, *amodal* representation that predicts the motion of the invisible stimulus. The shading of the neurons in FIGURE 4 shows the neuron activations during an occlusion–disocclusion sequence. This network does not contain Class III neurons; instead, the Class I neurons respond to the unpredicted appearance of moving objects.

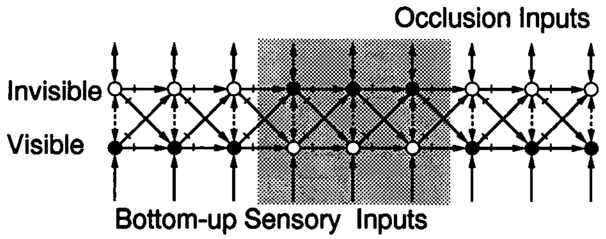

Figure 4: Network architecture after learning. The learning procedure causes the representation of each trajectory to split into two parallel opponent channels. The Visible and Invisible channel pair for a single trajectory are shown. Solid arrows indicate excitatory connections. Hash marks indicate time-delayed connections. Broken arrows indicate inhibitory connections. The network's responses to an occlusion–disocclusion sequence are shown. A visual object (not shown) moves rightward along the trajectory. The object is represented successively at each moment by neuron activations in the Visible channel. (Solid circles represent active neurons; open circles represent inactive neurons.) When the object reaches the region indicated by gray shading, it disappears behind an occluder. The network responds by successively activating neurons in the Invisible channel. These neuron activations amodally represent the network's inference that the object continues to exist and continues to move while it is invisible. When the object emerges from behind the occluder (end of gray shading), it is again represented by activation in the Visible channel. Because the Visible and Invisible channels are distinct, the neurons in the Invisible channel can also learn bidirectional associations (reciprocal excitatory connections) with other neurons whose activation is correlated with the presence of an invisible object. If any such neurons exist, we can call them occlusion indicators.

The network generates many predictions at every moment; each prediction's weight is determined partially by its learned statistical likelihood. Each prediction generated by the system has one of three possible outcomes: (1) an object appears at the predicted location; (2) an object does not appear at the predicted location but does appear at another predicted location; or (3) no object appears at a predicted location. The disinhibition rules ensure that one of these three outcomes is attributed to every prediction.

4 Learning Relative Depth from Occlusion Events

As we have described, the network develops Off channels that represent occluded objects. The activation of neurons in the Off channels is highly likely to be correlated with the activation of other neurons elsewhere in the visual system, specifically neurons whose activation indicates the presence of occluders. Simple Hebbian-type learning will let such occlusion-indicator neurons gradually establish excitatory connections to the Off channel neurons, and vice versa.

After such reciprocal excitatory connections have been learned, the activation of occlusion-indicator neurons at a given spatial position tends to cause the network to favor the Off channel in its predictions – i.e., to predict that a moving object will be invisible at that position. Thus, the network learns to use occlusion information to generate better predictions of the visibility/invisibility of objects.

Conversely, the activation of Off channel neurons causes the occlusion-indicator neurons to receive excitation. The disappearance of an object excites the representation of an occluder at that location. If the representation of the occluder was not previously activated, then the excitation from the Off channel may even be strong enough to activate it alone. Thus, the disappearance of moving visual objects constitutes evidence for the presence of an inferred occluder.

We believe that our model of depth perception is consistent with Grossberg's (1993) figure-ground detection network, and we believe that ours can lead to a self-organizing network that has similar figure-ground detection behavior.

5 Conclusions: Depth, On/Off Channels, Disinhibition, and Learning

Two results that have emerged from this research suggest how visual systems may learn to represent visual depth information. First, a visual system can learn a nonmetric representation of the depth relations arising from occlusion events. The existence of separate channels for representing modal and amodal information provides a substrate upon which the network can learn associations between the disappearance of moving objects and the presence of occluders.

Second, parallel opponent On and Off channels that represent both modal and amodal stimuli can also be learned through the same process. These channels let visual systems represent exceptions to some predictive rules – e.g., to represent an occlusion event as an exception (the unexpected failure to appear) to a predicted event (the appearance of an object at a certain location). The On and Off channels thus improve the ability of visual systems to model, predict, and represent the visual appearance of the world.

Acknowledgements

Supported in part by the Office of Naval Research (Cognitive and Neural Sciences, N00014-93-1-0130), the National Eye Institute (EY09669), a UNC-CH Junior Faculty Development Award, an ORAU Junior Faculty Enhancement Award from Oak Ridge Associated Universities, the University of Minnesota Center for Research in Learning, Perception, and Cognition, the National Institute of Child Health and Human Development (HD-07151), and the Minnesota Supercomputer Institute.

We thank Stephen Aylward, Eliza Graves, Robert Hubbard, and Al Nigrin for many valuable discussions.

References

Albus JS (1991) Outline for a theory of intelligence. *IEEE Transactions on Systems, Man, and Cybernetics* 21:473–509.

Anstis S, Ramachandran VS (1986) Entrained path deflection in apparent motion. *Vision Research* 26:1731–1739.

Bodnarenko SR, Chalupa LM (1993) Stratification of On and Off ganglion cell dendrites depends on glutamate-mediated afferent activity in the developing retina. *Nature* 364:144–146.

Frost BJ (1993) Time to collision sensitive neurons in nucleus rotundus of pigeons. Conference talk, *Workshop on Binocular Stereopsis and Optic Flow*, York Univ., Toronto, Canada.

Grossberg S (1993) A solution of the figure-ground problem for biological vision. *Neural Networks* 6:463–483.

Kaplan GA (1969) Kinetic disruption of optical texture: The perception of depth at an edge. *Perception & Psychophysics* 6:193–198.

Marshall JA (1989) Self-organizing neural network architectures for computing visual depth from motion parallax. *Proceedings of the International Joint Conference on Neural Networks*, Washington DC, II:227–234.

Marshall JA (1990) Self-organizing neural networks for perception of visual motion. *Neural Networks* 3:45–74.

Marshall JA (1991) Challenges of vision theory: Self-organization of neural mechanisms for stable steering of object-grouping data in visual motion perception. *Stochastic and Neural Methods in Signal Processing, Image Processing, and Computer Vision*, S.-S. Chen, Ed., Proceedings of the SPIE 1569, San Diego, CA, 200–215.

Marshall JA (1992) Unsupervised learning of contextual constraints in neural networks for simultaneous visual processing of multiple objects. *Neural and Stochastic Methods in Image and Signal Processing*, S.-S. Chen, Ed., Proceedings of the SPIE 1766, San Diego, CA, 84–93.

Marshall JA (1993) Adaptive perceptual pattern recognition by self-organizing neural networks: Context, uncertainty, multiplicity, and scale. Submitted for publication, 46 pp.

Nakayama K, Shimojo S (1992) Experiencing and perceiving visual surfaces. *Science* 257:1357–1363.

Nakayama K, Shimojo S, Silverman GH (1989) Stereoscopic depth: Its relation to image segmentation, grouping, and the recognition of occluded objects. *Perception* 18:55–68.

Ramachandran VS, Inada V, Kiama G (1986) Perception of illusory occlusion in apparent motion. *Vision Research* 26:1741–1749.

Shimojo S, Silverman GH, Nakayama K (1988) An occlusion-related mechanism of depth perception based on motion and interocular sequence. *Nature* 333:265–268.

Shimojo S, Silverman GH, Nakayama K (1989) Occlusion and the solution to the aperture problem for motion. *Vision Research* 29:619–626.

Yonas A, Craton LG, Thompson WB (1987) Relative motion: Kinetic information for the order of depth at an edge. *Perception & Psychophysics* 41:53–59.