

Does Visual Development Aid Visual Learning?

Melissa Dominguez

Department of Computer Science

University of Rochester

Rochester, NY 14627

Robert A. Jacobs

Department of Brain and Cognitive Sciences

University of Rochester

Rochester, NY 14627

January 2002

This work was supported by NSF Graduate Fellowship DGE9616170 and by NIH research grant R01-EY13149.

1 Introduction

Relative to adults, human infants are born with limited perceptual, motor, linguistic, and cognitive abilities. There are at least two perspectives within the field of developmental psychology regarding these limitations. The older and more popular view is that these limitations are barriers which must be overcome in order for a child to achieve adult function (Piaget, 1952). That is, they are immaturities or deficiencies which serve no positive purpose. A newer view is that these apparent inadequacies are in fact helpful, perhaps necessary, stages in development. According to this theory, limited mental abilities reflect simple neural representations which are useful “stepping stones” or “building blocks” for the subsequent development of more complex representations (Turkewitz and Kenney, 1982).

The development of biological nervous systems is sometimes characterized in terms of this newer view. Greenough, Black, and Wallace (1987) hypothesized that asynchrony in brain development serves the useful function of “stage setting.” The developmental schedule for the maturation of different brain regions is staggered such that neural systems that develop relatively early provide a suitable framework for the development of later, experience-sensitive systems. The notion of stage-setting can be generalized by considering the fact that neural events can be ordered either in time, such as temporally earlier versus later events, or in space, as in events at earlier versus later regions in a hierarchy of brain regions (e.g., lateral geniculate nucleus versus primary visual cortex).

Neuroscientific evidence consistent with the stage-setting hypothesis comes from Harwerth, Smith, Duncan, Crawford, and von Noorden (1986). These investigators performed behavioral studies of sensitive periods for visual development in monkeys. Their results suggest that these sensitive periods are organized into a hierarchy in which early visual

functions requiring information processing in the peripheral portions of the visual system have shorter sensitive periods than higher-level functions requiring more central neural processing. A second example is provided by the work of Shatz (1996). Within the lateral geniculate nucleus (LGN) of adult mammals, retinal ganglion cell axons from one eye are segregated from those arising from the other eye to form a series of alternating eye-specific layers. These layers are not present initially in development. Moreover, they form during a period in which vision is not possible. Shatz argued that the development of eye-specific layers is characterized by at least two important events. First, retinal ganglion cells spontaneously show waves of activity that sweep across the retina such that activity at nearby cells is more highly correlated than at distant cells. Second, LGN cells use a Hebb-style adaptation mechanism that sorts connections based on local correlations of activity in order to form eye-specific layers. If so, then this is a clear example in which a developmental event at an earlier visual region (spontaneous waves of activity at the retina) sets the stage for an event at a later visual region (Hebb-style adaptation at the LGN).

Newport (1990) hypothesized that children use a stage-setting strategy when attempting to learn a language. Human languages are componential systems in which large linguistic structures are formed by systematically combining smaller components. According to Newport's (1990) "Less is More" hypothesis, the limited attentional and memorial abilities of children are useful when learning a language because they help children segment and identify the components that comprise the language. An implementation of this general idea was studied by Elman (1993). He showed that a recurrent neural network whose memory capacity was initially limited but then gradually increased during the course of training learned aspects of a grammar better than a network whose memory capacity was never limited; i.e. the second network's memory capacity was always equal to that of the

first network at the end of training. Elman argued that this outcome supports the idea that “starting small” is a developmental property that is important to the subsequent acquisition of complex mental abilities. Rohde and Plaut (1999), however, were unable to replicate Elman’s simulation results, so it is difficult to know how to interpret these results.

This chapter considers the hypothesis that systems learning aspects of visual perception may benefit from the use of suitably designed developmental progressions during training. Relative to adults, human infants are born with poor visual acuity (acuity refers to the amount of detail that can be detected or discriminated in a visual image). The acuity of newborns is about 20/400 when measured on the Snellen scale whereas that of normal adults is 20/20. Their acuity improves approximately linearly from these low levels at birth to near adult levels by about 8 months of age (Norcia and Tyler, 1985). At the same time, infants are acquiring other visual abilities such as the ability to detect binocular disparities (binocular disparities refer to differences in left and right visual images due to the fact that our eyes are offset from each other; these disparities carry information about the three-dimensional depth and shape of objects in the visual environment). Sensitivity to binocular disparities appears at around 4 months of age (Atkinson and Braddick, 1976; Fox, Aslin, Shea, and Dumais, 1980; Held, Birch, and Gwiazda, 1980). Is there a functional relationship between the development of visual acuity and the development of binocular disparity sensitivities? We speculate that the developments of acuity and binocular disparity sensitivity may be related in the sense that, counter-intuitively, poor acuity at an early age aids the acquisition of disparity sensitivity later in life.

Unfortunately, essentially nothing is known about the psychological processes or mental representations underlying the detection and use of binocular disparities and, thus, nothing is known about how these underlying mental processes and representations change

during development. Consequently, our attempt to understand the development of binocular disparity sensitivities emphasizes neuroscientific variables, not psychological variables. Many simple and complex cells in primary visual cortex of primates, though not all, are sensitive to binocular disparities, and a great deal is known about these cells' receptive field properties from both neuroscientific and computational viewpoints.

We report the results of simulations in which three different systems were exposed to pairs of visual images, where each pair depicted an object against a background or a fronto-parallel surface. The systems were trained to detect the binocular disparity of the object or of the surface, meaning the shift of the object or surface from the right-eye image to the left-eye image. Two of the systems are developmental models in the sense that the nature of their input changed during the course of training. The third system is a non-developmental model; its input remained constant during the course of training. The inputs to the systems were left and right retinal images filtered with binocular energy filters tuned to various spatial frequencies. Binocular energy filters are a computational tool for modeling the binocular sensitivities of simple and complex cells in primary visual cortex of primates. The training of the first system, referred to as the *coarse-scale-to-multiscale model* (or model C2M), included a developmental sequence such that the system was exposed only to low spatial frequency information at the start of training (i.e. coarse-scale disparity features), and information at higher spatial frequencies (mid-scale and fine-scale disparity features) was added to its input as training progressed. The training of the second system, referred to as the *fine-scale-to-multiscale model* (or model F2M), included an analogous developmental sequence with spatial frequency information added in the reverse order. This system received high spatial frequency information at the start of training (fine-scale disparity features); information at lower spatial frequencies (mid-scale and coarse-scale disparity features) was added as training progressed. The third

system, referred to as the *non-developmental model* (or model ND), was not trained using a developmental sequence; it received information at all spatial frequencies throughout the training period.

When comparing the two developmental models with the non-developmental model, there are at least two reasonable predictions that one could make about the simulation results. One prediction is that the non-developmental model should outperform the developmental models. The non-developmental model received all input information throughout all stages of training, whereas the developmental models were deprived of portions of the input at certain training stages. If more information is better than less information, then the non-developmental model ought to perform best. This would be consistent with the traditional view of human infant development, that perceptual immaturities are barriers to be overcome.

An alternative prediction, consistent with the general approach of the “less is more” hypothesis described above, is that the developmental models would show the best performance. If it is believed that too much information could lead a learning system in its early stages of training to form poor internal representations, then the developmental models ought to have an advantage. Relative to the developmental models, the non-developmental model had a greater number of inputs and, thus, a greater number of modifiable weights during the early stages of training. Because learning in neural networks is a search in weight space, the non-developmental model needed to perform an unconstrained search in a high-dimensional weight space. Unconstrained searches frequently lead to the acquisition of poor representations. In contrast, the developmental models initially had fewer inputs and, thus, fewer modifiable weights. During early stages of training, their searches in weight space were comparatively constrained. If the constraints were appropriate,

this should have facilitated the acquisition of useful representations by the developmental models.

When comparing the performances of the two developmental models, there are at least three reasonable predictions that one could make about the simulation results. One prediction is that the coarse-scale-to-multiscale model should perform best. A motivation for this prediction comes from the field of computer vision. Consider the task of aligning two images of a scene where the images differ due to a small horizontal offset in their viewpoints. Roughly, this is known as the stereo correspondence problem. If this were something that you had never tried to do before, you might try to align fine details of each image. For example, you might pick a white dot in one image and repeatedly try aligning it with dots in the other image. If the images are highly textured, such an approach would be inefficient as there is an intractable number of potential alignments that might need to be checked before the two images are properly aligned (see Panel A of Figure 1). If, however, the images are blurred, the fine details which were the source of so much confusion would be removed, leaving a smaller number of larger features in each image. The problem of finding a good alignment is now significantly easier because there is a smaller number of potential alignments (see Panel B of Figure 1). After you have gained skill at aligning blurred images, you would then have a reliable foundation that you could use to learn to align clearer images. You might attempt to match the larger image features first. Subsequent analysis of fine details would seek to remove ambiguities that arise when aligning large features, instead of being a starting point for locating correspondences.

This style of processing is commonplace in the computer vision literature. For example, systems by Marr and Poggio (1979), Quam (1986), and Barnard (1987) initially search for correspondences within a pair of low resolution images. Low resolution images are used initially because these images contain fewer image features, larger image

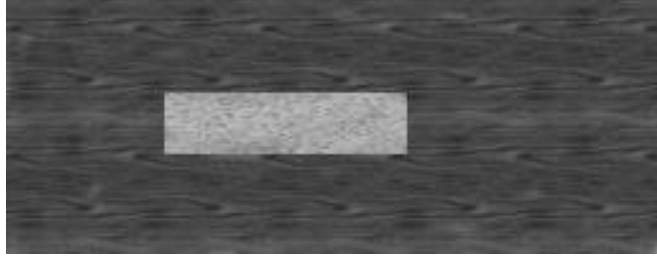
A**B**

Figure 1: A: A textured object and background. B: The image from Panel A blurred so as to remove the fine details.

features, and image features that are relatively robust to noise. Next, these systems refine their estimates of corresponding image features using information from one or more higher-resolution pairs of images. The usefulness of a coarse-to-fine processing strategy when searching for stereo correspondences suggests that a coarse-scale-to-multiscale developmental strategy might be useful for learning to detect binocular disparities. Before adopting this hypothesis, however, it is important to keep in mind the differences between these two strategies. Computer vision researchers use a coarse-to-fine strategy when searching for correspondences in individual pairs of images. In contrast, we used the coarse-scale-to-multiscale developmental sequence while training a learning system to detect binocular disparities using many pairs of images. It is not obvious that lessons from one situation can be applied to the other situation. Nonetheless, it may be the case that the use of a coarse-scale-to-multiscale developmental sequence biases a learning system so that it initially develops an approximate solution based on coarse-scale information and

then subsequently learns to refine this solution using fine-scale information. If so, then the style of processing learned by this system at the end of training may resemble that of a non-adaptive system that is designed to use a coarse-to-fine processing style.

If we assume that human intelligence provides a guide to creating machine intelligence, then a second motivation for the prediction that the coarse-scale-to-multiscale developmental model should perform best is the fact that human infants show a related developmental progression. Visual acuity is often measured using a grating, which is a visual pattern whose luminance values are sinusoidally modulated. Acuity is characterized by the highest-frequency grating which is distinguishable from a solid gray pattern. Using this method, it has been found that newborns' visual acuity is extremely poor. Whereas adults with normal vision (so-called 20/20 vision) can discriminate approximately 30 cycles per degree of arc, newborns can only discriminate 1–2 cycles per degree giving them a visual acuity of about 20/400. Acuity improves approximately linearly from these low levels at birth to near adult levels by around 8 months of age (Norcia and Tyler, 1985). Importantly for our purposes, infants are acquiring other visual abilities during this time period; in particular, sensitivity to binocular disparities appears at around 4 months of age (Atkinson and Braddick, 1976; Fox, Aslin, Shea, and Dumais, 1980; Held, Birch, and Gwiazda, 1980; Petrig, Julesz, Kropfl, Baumgartner, and Anliker, 1981). We speculate that the developments of visual acuity and binocular disparity sensitivity may be related in the sense that poor acuity at an early age aids in the acquisition of disparity sensitivity later in life.

An alternative prediction is that the fine-scale-to-multiscale developmental model should perform best. A motivation for this prediction is the fact that computer vision researchers often find it easier to solve the stereo correspondence problem by first extracting edge information from left and right images, a form of high-frequency band-pass

filtering, and then searching for a good alignment of the images based on this information. This strategy is useful because edges can be sparse, large, and robust to noise relative to other image features. Analogous to computer vision systems' initial use of high-frequency information, the fine-scale-to-multiscale model is initially trained solely with information extracted from high-frequency band-pass filters. If we are willing to assume that computer vision methods for finding stereo correspondences may provide lessons for how learning systems can learn to detect binocular disparities, as discussed above, then we might predict that model F2M has an advantage.

A second motivation for the prediction that model F2M should perform best is the seemingly counter-intuitive result that neural networks trained with input patterns that have been corrupted by noise frequently show better generalization than equivalent networks trained with input patterns that have not been corrupted (Sietsma and Dow, 1991). Training with noisy inputs has been shown analytically to be equivalent to a form of regularization (Bishop, 1995; Matsuoka, 1992; Webb, 1994). In other words, the learning process of networks trained with noisy inputs is more constrained than that of similar networks whose inputs are not corrupted by noise. If the retinal images contain noise, and if the coarse-scale-to-multiscale model tends to filter out the noise during early stages of training, then this model will not obtain the benefits of training with noisy inputs. We would, therefore, expect the fine-scale-to-multiscale model to perform best.

Lastly, another possible prediction is that the coarse-scale-to-multiscale model and the fine-scale-to-multiscale model perform about equally well. If the relative advantages of model C2M (being able to use fewer image features, larger image features, and more noise-resistant image features at the start of training) and the relative advantages of model F2M (initial exposure to high-frequency band-pass information, being able to obtain the benefits of training with noisy inputs) are roughly balanced, then neither model would

be expected to show superior performance. Furthermore, we outlined above the logic of computer vision researchers who advocate a coarse-to-fine processing strategy when analyzing stereo correspondences, and we tentatively speculated that the usefulness of this strategy suggests that the seemingly related coarse-scale-to-multiscale developmental strategy might be useful for learning to detect binocular disparities. If, however, we again make the assumption that human intelligence provides a guide to creating machine intelligence, then it is worth noting that psychologists have discovered that humans often do not use a coarse-to-fine strategy when analyzing stereo correspondences. Mallot, Gillner, and Arndt (1996) found that unambiguous information at a coarse scale is not always used by observers to disambiguate finer scale information, and found that observers can use unambiguous fine-scale information to disambiguate coarse-scale information, meaning that observers are using a fine-to-coarse processing strategy in these circumstances. Related findings have been reported by several other researchers (e.g., McKee and Mitchison, 1988; Mowforth, Mayhew, and Frisby, 1981; Smallman, 1995). Because human observers neither exclusively use a coarse-to-fine strategy nor a fine-to-coarse strategy, we might not expect the exclusive use of a coarse-scale-to-multiscale developmental strategy or the exclusive use of a fine-scale-to-multiscale developmental strategy to yield a relative performance advantage.

This chapter reports the results of computer simulations comparing the learning performances of developmental and non-developmental models on the task of estimating binocular disparities of objects against backgrounds or of fronto-parallel surfaces in novel pairs of images. Three results are highlighted. First, developmental models C2M and F2M consistently outperformed the non-developmental model ND. Second, models C2M and F2M outperformed a random developmental model that was also trained in stages, but the inputs to this model were randomly selected at each stage. Finally, models C2M

and F2M outperformed developmental models whose inputs did not progress in an orderly fashion from one resolution scale to a neighboring scale during the course of development. Additional details and analyses regarding these results can be found in Dominguez and Jacobs (2002). The fact that models C2M and F2M outperformed the non-developmental model is important because this demonstrates that models that undergo a developmental maturation can acquire a more advanced perceptual ability than one that does not. The fact that models C2M and F2M outperformed the random developmental model is important because this demonstrates that not all developmental sequences can be expected to provide performance benefits. To the contrary, only sequences whose characteristics are matched to the task should lead to superior performance. Our findings suggest that the most successful models are those that are exposed to visual inputs at a single scale early in training, and for which the resolution of their inputs progresses in an orderly fashion from one scale to a neighboring scale during the course of training. On the basis of these results, we conclude that developmental sequences can be useful to systems learning to detect binocular disparities, and that the general idea that visual development can aid visual learning is a viable hypothesis in need of future study.

2 Developmental and Non-Developmental Models

The structure of the developmental and non-developmental models is illustrated in Figure 2. This structure is based on a similar architecture studied by Gray, Pouget, Zemel, Nowlan, and Sejnowski (1998). The retinal layer consisted of two one-dimensional arrays 62 pixels in length for the left and right eye images. Each retina was treated as if it were shaped like a circle; the leftmost and rightmost pixels were regarded as neighbors. This wraparound of the left and right edges was done to avoid edge artifacts. Although one-dimensional retinas are a simplification, their use is justified by the fact that the

models were only concerned with horizontal disparities as these are the ones that provide information about the three-dimensional configuration of the visual environment (vertical disparities provide information about viewing distance and angle of gaze, but not about the three-dimensional nature of the environment). The retinal inputs were filtered using binocular energy filters.

Based on neurophysiological studies, Ohzawa, DeAngelis, and Freeman (1990) proposed binocular energy filters as a way of modeling the binocular sensitivities of simple and complex cells in primary visual cortex. These filters are a modification of motion energy filters proposed by Adelson and Bergen (1985). Intuitively, these filters may be understood as follows. When tested with light patterns presented to a single eye, simple cells are often found to have a receptive-field profile that is accurately modeled by a mathematical function known as a Gabor function. Consider the Gabor function shown in the upper-left corner of Figure 3. The horizontal axis in this graph gives retinal position (along an axis perpendicular to the neuron's preferred orientation), and the vertical axis represents sensitivity to a luminance increment such that upward and downward deflections of the curve correspond to ON and OFF subregions of the neuron's receptive field, respectively. Based on the receptive-field profile in the upper-left corner of this figure, the simple cell would tend to increase its activity in response to a luminance increment in the center of its receptive field, and to decrease its activity in response to luminance increments away from its receptive-field center.

We need to consider both its left-eye and right-eye receptive-field profiles in order to understand this cell's sensitivity to binocular disparities. The top-row of Figure 3 illustrates these profiles for a hypothetical cell. The left-eye and right-eye profiles are identical in shape in this example. Note that this cell will have its largest response when there are luminance increments in the centers of both eyes' receptive fields. Now consider

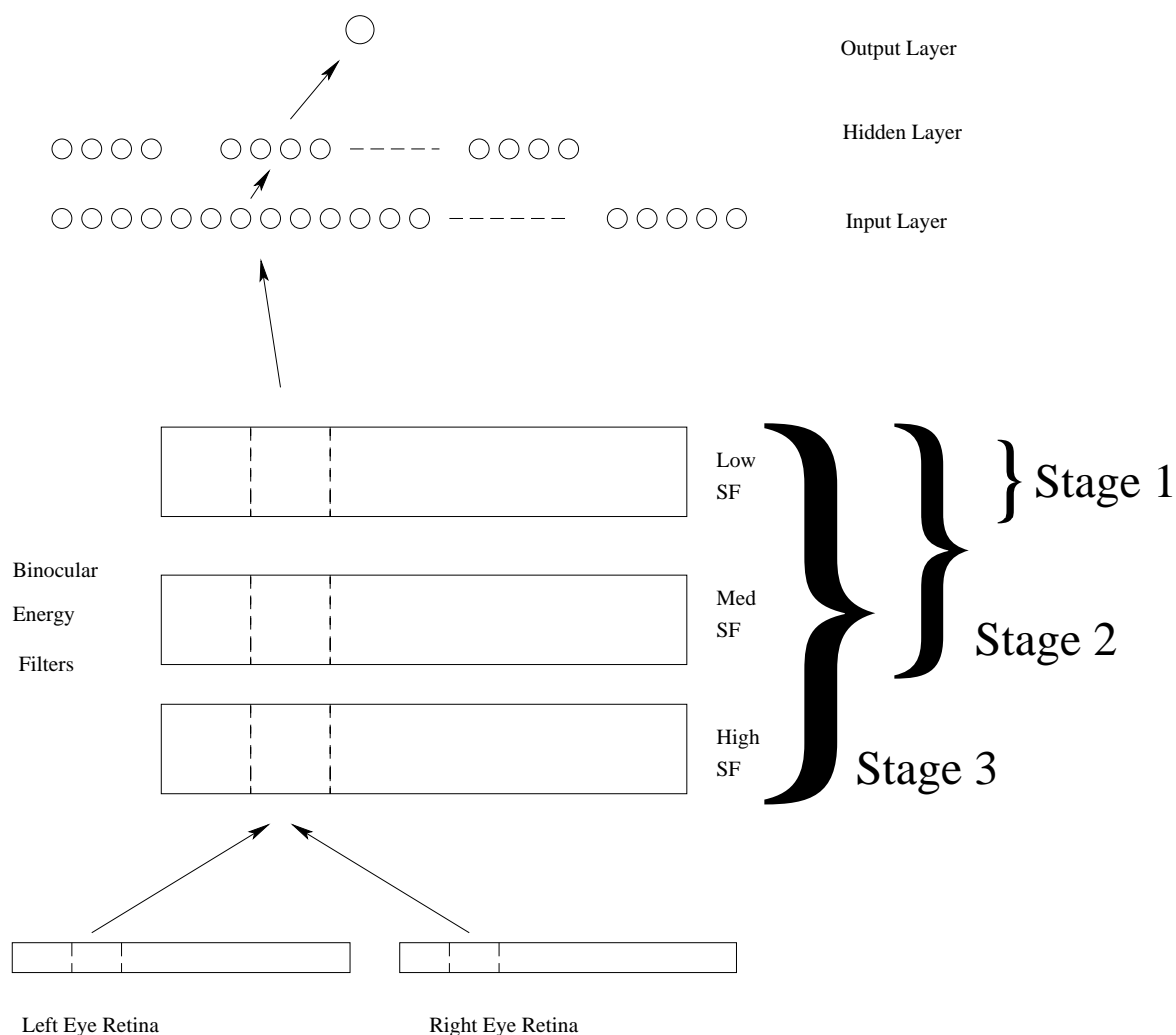


Figure 2: All models shared a common structure. The first portion of this structure is the left and right retinal images. These images are then filtered by binocular energy filters tuned to low, medium, and high spatial frequencies which simulate the binocular sensitivities of simple and complex cells in primary visual cortex. The outputs of these filters are the inputs to an artificial neural network that is trained to estimate the disparity present in the images. The model illustrated here is the coarse-scale-to-multiscale model in which low spatial frequency information was received during early stages of training, and information at higher frequencies was added as training progressed.

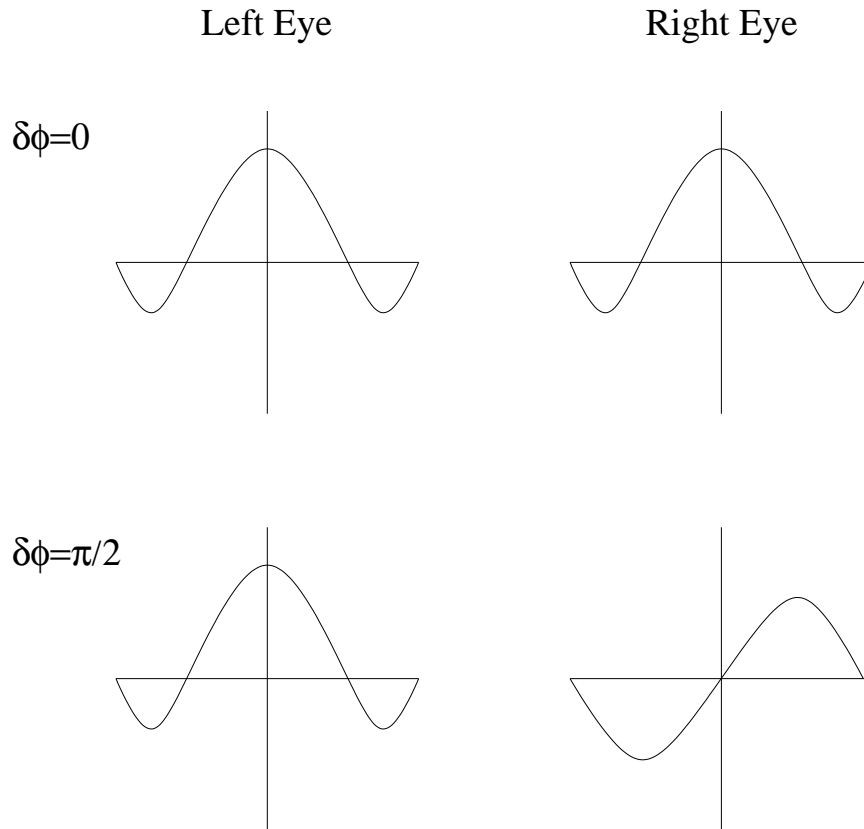


Figure 3: Illustrations of left-eye and right-eye receptive-field profiles. The simple cell whose profiles are illustrated in the top row is sensitive to zero disparities (no shift between left-eye and right-eye images), whereas the cell whose profiles are illustrated in the bottom row is sensitive to non-zero disparities (a shift of a certain direction and magnitude between left-eye and right-eye images).

the receptive-field profiles illustrated in the bottom row of Figure 3. The left-eye and right-eye profiles have different shapes in this case. This cell will have its largest response when there is a luminance increment in the center of its left-eye receptive field and an increment towards the right of its right-eye receptive field. In other words, this cell will have its largest response when a luminance increment in the left-eye image has shifted to the right in the right-eye image. Whereas the neuron illustrated in the top row is sensitive to zero disparities (no shift between left-eye and right-eye images), the neuron illustrated in the bottom row is sensitive to a non-zero disparity (a shift of a certain direction and magnitude between left-eye and right-eye images).

Although not illustrated in Figure 3, complex cells are hypothesized to sum the outputs of appropriate sets of simple cells so that their responses are phase-invariant, meaning that the responses do not depend on the exact position of a disparity within the cell's receptive field. The output of a binocular energy filter is meant to model the output of a complex cell. Our simulations used filters with different phase offsets between the left-eye and right-eye receptive-field profiles (these different filters were sensitive to disparities of different sizes), and different spatial frequency scales (these different filters were sensitive to disparity features at different image resolutions, ranging from coarse, or low-resolution, disparity features to fine, or high-resolution, disparity features). When considering the simulation results discussed below, it is important to keep in mind that complex cells (and binocular energy filters) indicate the disparities present in (local) image patches, and do not indicate the (global) disparity of an object against a background or of a fronto-parallel surface.

Mathematically, binocular energy filters are characterized as follows. A simple cell receives input from a pair of subunits, one for each retina. The receptive field profiles of the subunits can be described mathematically as Gabor functions:

$$g_L(x, \phi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \sin(2\pi\omega x + \phi) \quad (1)$$

$$g_R(x, \phi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \sin(2\pi\omega x + \phi + \delta\phi). \quad (2)$$

Each function is a sinusoid multiplied by a Gaussian envelope. The parameter x is the distance to the center of the Gaussian, σ^2 is the variance of the Gaussian, ω is the frequency of the sinusoid, and ϕ and $\delta\phi$ are referred to as the base phase and phase offset of the sinusoid. The Gabor functions associated with the left and right retinal subunits differ in that the phase of one is offset from the phase of the other (a non-zero phase offset accounts for why left-eye and right-eye receptive-field profiles have a different shape). The output of a simple cell is formed in two stages: first, the convolution of the left retinal image with the left subunit Gabor is added to the convolution of the right retinal image with the right subunit Gabor; next, this sum is half-wave rectified and squared (a negative sum is mapped to zero; a positive sum is mapped to its square). The magnitude of a simple cell's output is related to the presence of a binocular disparity of a particular size in the retinal input. Simple cells formed from subunits with different phase offsets are sensitive to disparities of different sizes (Fleet, Wagner, and Heeger, 1996; Qian, 1994). The output of a complex cell is the sum of the outputs of four simple cells. Because the base phases of these simple cells form quadrature pairs (the base phases are 0 , $\pi/2$, π , and $3\pi/2$), the complex cell's output is relatively insensitive to the exact position of a disparity within its receptive field.

In our simulations, there were 35 receptive-field locations which received input from overlapping regions of the retina. At each of these locations, there were 30 complex cells corresponding to 3 spatial frequencies and 10 phase offsets at each frequency. The three spatial frequencies were each separated by an octave: 0.25, 0.125, and 0.0625 cycles per pixel. The standard deviations of the Gabor functions were set to be inversely proportional to the frequency: 1.0 for 0.25 cycles/pixel, 2.0 for 0.125 cycles/pixel, and 4.0 for 0.0625 cycles/pixel. The ten phase offsets were equally spaced over a range from 0 to $\pi/2$. The outputs of the complex cells were normalized using a softmax nonlinearity:

$$\hat{E}_i(x) = \frac{e^{E_i(x)/\tau}}{\sum_j e^{E_j(x)/\tau}} \quad (3)$$

where $E_i(x)$ was the initial output of the complex cell, $\hat{E}_i(x)$ was the normalized output, τ is a scaling parameter known as a temperature parameter (its value was set to 0.25), and j indexed the 10 complex cells with different phase offsets at a receptive-field location within a single frequency band. As a result of this normalization, complex cells tended to respond to relative contrast in an image, rather than absolute contrast.

The normalized outputs of the complex cells were the inputs to an artificial neural network. As illustrated in Figure 4, the network had 1050 input units (the complex cells had 35 receptive field locations and there were 30 cells at each location). The hidden layer of the network contained 32 units which were organized into 8 groups of 4 units each. The connectivity to the hidden units was set so that each group had a limited receptive field; a group of hidden units received inputs from seven receptive field locations at the complex cell level. The hidden units used a logistic activation function. The output layer consisted of a single linear unit; this unit's output was an estimate of the object or surface disparity present in the right and left retinal images.

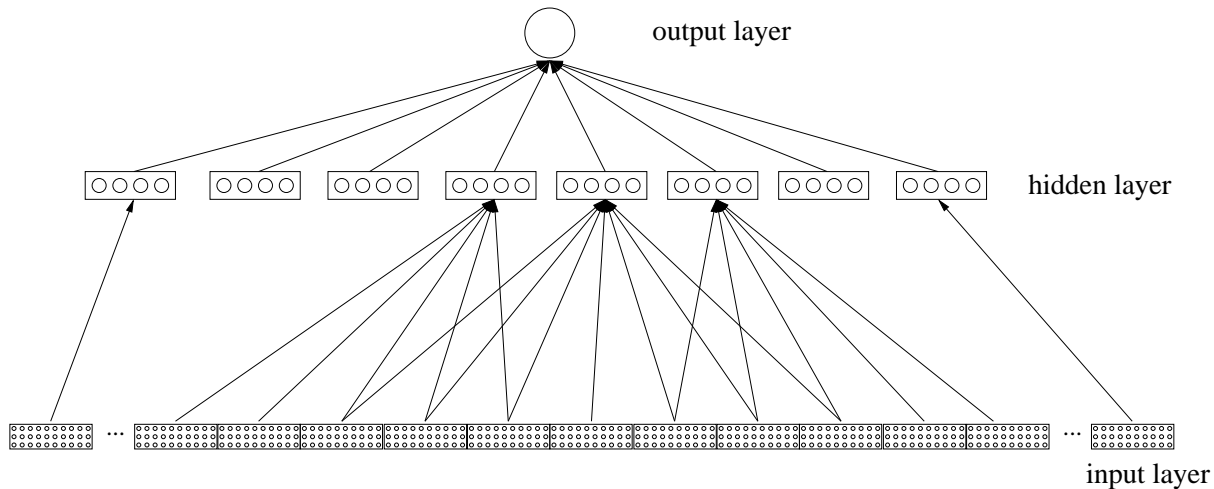


Figure 4: The structure of the artificial neural network portion of the developmental and non-developmental models.

The weights of an artificial neural network were initialized to small random values, and were adjusted during the course of training to minimize a sum of squared error cost function using a conjugate gradient optimization procedure (Press, Teukolsky, Vetterling, and Flannery, 1992). This procedure was used because it tends to converge quickly and because it has no free parameters (e.g., no learning rate or momentum parameters). Weight sharing was implemented at the hidden unit level so that corresponding units within each group of hidden units had the same incoming and outgoing weight values, and so that a hidden unit had the same set of weight values from each receptive field location at the complex unit level. This provided the network with a degree of translation invariance, and also dramatically decreased the number of modifiable weight values in the network. It therefore decreased the number of data items needed to train the network, and the amount of time needed to train the network.

Models were trained and tested using separate sets of training and test data items. Training sets contained 250 randomly generated data items; test sets contained 122 data items that were generated so as to uniformly cover the range of possible binocular dis-

parities. Training was terminated after 35 iterations through the training set in order to minimize over-fitting of the training data. The results reported below are based on the data items from the test set.

Model C2M was trained using a coarse-scale-to-multiscale developmental sequence. This was implemented as follows. The training period was divided into three stages where the first and second stages were each 10 iterations and the third stage was 15 iterations. During the first stage, the neural network portion of the model only received the outputs of complex cells tuned to low spatial frequencies (low-resolution disparity features). The outputs of the other complex cells were set to zero. During the second stage, the network received the outputs of complex cells tuned to low and medium spatial frequencies (low- and mid-resolution disparity features); it received the outputs of all complex cells during the third stage (low-, mid-, and high-resolution disparity features). The training of model F2M was identical to that of model C2M except that its training used a fine-scale-to-multiscale developmental sequence. During the first stage of training, its network received the outputs of complex cells tuned to high spatial frequencies. This network received the outputs of complex cells tuned to high and medium frequencies during the second stage, and received the outputs of all complex cells during the third stage. In contrast, the training period for the non-developmental model was not divided into separate stages; its neural network received the outputs of all complex cells throughout the training period.

3 Data Sets and Simulation Results

The performances of the three models were evaluated on three data sets. The data sets were based on related data sets used by Gray et al. (1998). In all cases the images were gray scale with luminance values between 0 and 1, and disparities with values between 0 and 3 pixels. Ten simulations of each model on each data set were conducted.

In the *solid object data set*, images consisted of a single light or dark object on a gray background. The object’s gray-scale value was either between 0.0 and 0.1 or between 0.9 and 1.0, whereas the gray-scale value of the background was always 0.5. The location of the object was randomly chosen to be a real-valued location on the retina. The object’s disparity was randomly chosen to be a real value between 0 and 3 pixels (i.e. in going from the right-eye image to the left-eye image the position of the object was shifted by 0-3 pixels). The object’s size was randomly chosen to be a real value between 10 and 25 pixels. Since the object’s size, location, and disparity were all real numbers, the ends of the object could fall at a real-valued location within a pixel. In these (common) cases the value of the partially covered pixel was interpolated between the gray-scale value of the object and that of the background in proportion to the amount of the pixel covered by the object and background. An example of a right and left image is shown in the top panel of Figure 5. Given the right and left images, the task of a model was to estimate the object’s disparity.

The leftmost graph of Figure 6 illustrates the results. The horizontal axis gives the model, and the vertical axis gives the root mean squared error (RMSE) at the end of training on the data items from the test set. On average, developmental model C2M had a 16.5% smaller generalization error than the non-developmental model (the difference between the mean error rates is statistically significant; $t = 3.77$, $p < 0.002$ using a two-tailed t-test). Developmental model F2M had a 12.2% smaller error than the non-developmental model ($t = 23.74$, $p < 0.001$). Clearly, the two developmental models outperformed the non-developmental model. A statistical comparison between the developmental models C2M and F2M shows that their performances were not significantly different.

The images in the second data set, referred to as the *noisy object data set*, were meant to resemble random-dot stereograms frequently used in behavioral experiments. Images

Solid Object



Noisy Object



Planar



Figure 5: Examples of right and left images (top and bottom rows in each panel) from the solid object, noisy object, and planar data sets. Top panel: in going from the top to the bottom images, the dark object is shifted to the right. Middle panel: the noisy object is towards the left edge of the images. In going from the top to the bottom images, the noisy object is shifted by two pixels to the right. Bottom panel: the entire top image is shifted by two pixels to the right in order to form the bottom image.

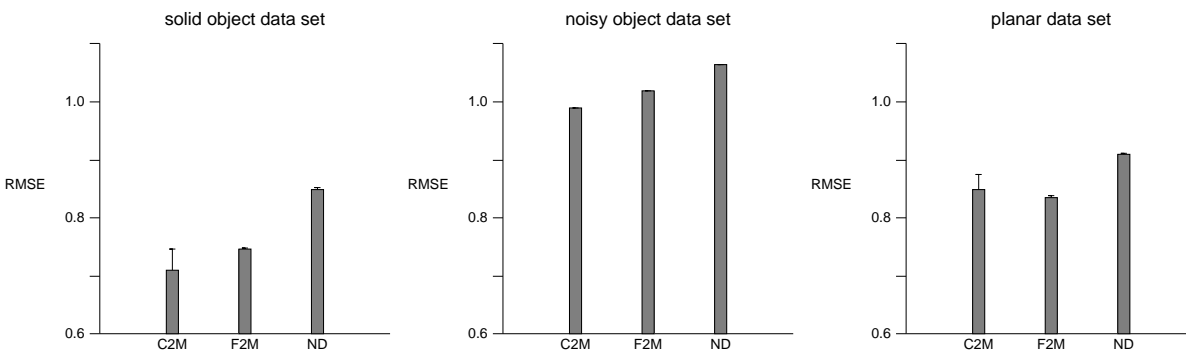


Figure 6: The three models' root mean squared errors (RMSE) on the test set data items after training on the three data sets (the error bars give the standard error of the mean).

contained a noisy object against a noisy background. The gray-scale values of the object pixels and the background pixels were set to random numbers between 0 and 1. The location of the object was randomly chosen to be a real-valued location on the retina. The object's size was randomly chosen to be a real value between 10 and 25 pixels. The object's disparity was a randomly chosen integer between 0 and 3 pixels. As before, the task was to map the left and right images to an estimate of the object's disparity. An example of a left and right image is shown in the middle panel of Figure 5.

The results are shown in the middle panel of Figure 6. As before, the developmental models consistently performed better than the non-developmental model. On average, model C2M had a 7.1% smaller generalization error than model ND, and model F2M had a 4.3% smaller error than model ND. Comparing the two developmental models, model C2M had a 2.85% smaller error than model F2M. (All the differences in the mean error rates are statistically significant; C2M versus ND: $t = 10.33$, $p < 0.001$; F2M versus ND: $t = 15.08$, $p < 0.001$; C2M versus F2M: $t = 3.68$, $p < 0.002$).

The last data set, the *planar data set*, was different from the first two data sets. Instead of an object in front of a background, the images depicted a fronto-parallel plane. The values of the left-image pixels were randomly chosen to be either 0 or 1. The right image was generated by applying an integer shift to the left image of 0, 1, 2, or 3 pixels. Given the left and right images, the task was to estimate the size of the shift. An example of a left and right image is shown in the bottom panel of Figure 5.

The results are shown in the rightmost graph of Figure 6. Again, the developmental models outperformed the non-developmental model. Model C2M had a 6.7% smaller generalization error than model ND ($t = 2.27$, $p < 0.05$), and model F2M had an 8.3% smaller error than model ND ($t = 16.84$, $p < 0.001$). The performances of models C2M and F2M were not statistically different.

Overall, the results show that the developmental models C2M and F2M consistently outperformed the non-developmental model ND. However, the results do not strongly suggest why the developmental models showed superior performance. In order to obtain a better understanding of this outcome, we conducted two additional sets of simulations.

A possible reason for why the developmental models showed superior performance is that they had relatively few inputs during the initial stages of training. In other words, it might be that the exact nature of the developmental progression is irrelevant. Instead, what matters is that the input size of a system starts small and gradually grows during the course of training. To check this possibility, we simulated a *random developmental model*, referred to as model RD. This system was similar to models C2M and F2M in the sense that its training included a developmental sequence. However, whereas the inputs received by models C2M and F2M at each developmental stage were organized by spatial frequency content, the inputs received by model RD at each stage were randomly selected. The collection of complex cells was randomly partitioned into three equal-sized subsets with the constraint that each subset included all phase offsets at all receptive field locations. During the first stage of training, the neural network portion of the model only received the outputs of the complex cells in the first random subset. It received the outputs of the cells in the first and second random subsets during the second stage of training, and received the outputs of all complex cells during the third stage.

The results of training and testing the developmental models C2M, F2M, and RD, and the non-developmental model ND on the solid object data set are shown in Figure 7. Model C2M had a 19.3% smaller generalization error than model RD ($t = 4.60$, $p < 0.001$), and model F2M had a 15.2% smaller error than the random developmental model ($t = 49.01$, $p < 0.001$). This result clearly indicates that the superior performances of models C2M and F2M were not due to limitations on their input sizes. Model RD had

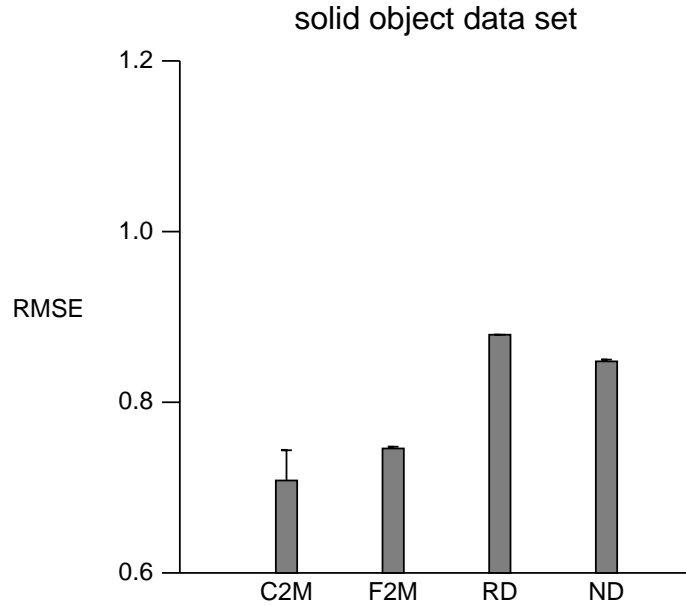


Figure 7: The root mean squared errors (RMSE) of developmental models C2M, F2M, and RD, and the developmental model ND on the test items from the solid object data set.

the same limitations but its performance was relatively poor. Instead it seems that there is something important about the exact nature of the developmental progressions that accounts for their superior performances.

We speculate that two important features of the developmental progressions of models C2M and F2M account for their superior performance. First, these models were exposed to visual inputs at a single scale or resolution early in training. Model C2M only received coarse-scale information at the start of training; model F2M only received fine-scale information. In contrast, models RD and ND received information at all spatial scales at all stages of training. We conjecture that it might be advantageous for a learning system to receive inputs at a single scale early in training because this allows the system to combine and compare input features without the need to compensate for the fact that these features could be at different spatial scales. Second, models C2M and F2M might be at an

advantage because the spatial scale of their inputs progressed in an orderly fashion from one scale to a neighboring scale. Consequently, when these models received inputs at a new spatial scale, this new scale was close to a scale with which the models were already familiar. If it is the case that this second feature of models C2M and F2M is important, then this leads to an interesting prediction. We predict that developmental models whose progressions do not proceed in an orderly manner from one scale to a neighboring scale ought to show poor performance.

To test this prediction, two additional models were created and tested on the solid object data set. These models had developmental stages based on spatial frequency content (like models C2M and F2M) but the addition of new frequency bands to their inputs at each stage did not proceed in an orderly manner. The first new model is referred to as model C-CF-CMF. It received the outputs of complex cells tuned to a low spatial frequency (coarse disparity features) early in training. In stage two it received the outputs of complex cells tuned to low and high frequencies (coarse and fine disparity features), and it received the outputs of cells tuned to low, medium, and high frequencies (coarse, medium, and fine disparity features) in stage three. The second new model, referred to as model F-CF-CMF, was similar to model C-CF-CMF but it started with high frequency information. That is, it received the outputs of complex cells tuned to a high spatial frequency early in training. It received the outputs of cells tuned to low and high frequencies in stage two, and the outputs of all complex cells in stage three.

Figure 8 shows the performances of models C2M, F2M, C-CF-CMF, and F-CF-CMF at the end of training on the solid object data set. In accord with our prediction, models C-CF-CMF and F-CF-CMF showed very poor performance. This data is consistent with the conjecture discussed above that it is advantageous to a developmental system for the

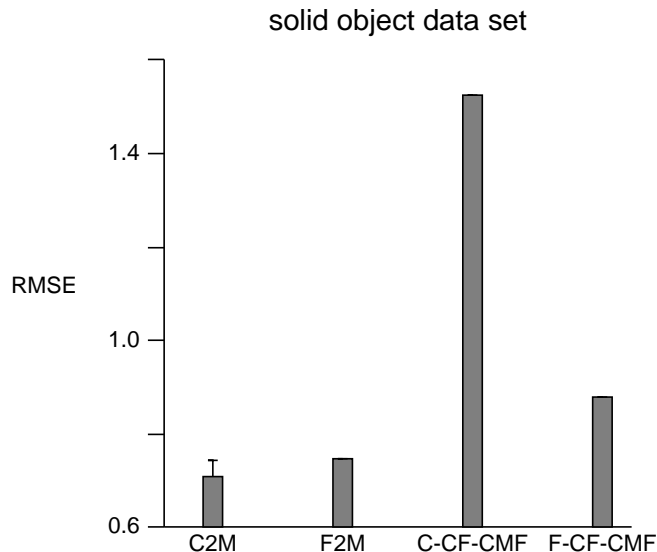


Figure 8: The root mean squared errors (RMSE) of models C2M, F2M, C-CF-CMF, and F-CF-CMF on the test items from the solid object data set.

spatial scale of its inputs to progress in an orderly fashion from one scale to a neighboring scale.

4 Concluding Remarks

We have considered the hypothesis that systems learning aspects of visual perception may benefit from the use of suitably designed developmental progressions during training. In particular, we have speculated that there is a functional relationship between the development of visual acuity and the development of binocular disparity sensitivities; namely, that poor acuity at an early age aids the acquisition of disparity sensitivity later in life. To evaluate this hypothesis, we simulated a number of developmental and non-developmental learning systems. Overall, the results suggest that the most successful systems at learning to detect the binocular disparities of solid or noisy objects against backgrounds or of planar surfaces are systems that are exposed to visual inputs at a

single scale early in training, and for which the resolution of their inputs progresses in an orderly fashion from one scale to a neighboring scale during the course of training. Given that human infants show a coarse-to-multiscale progression in their visual acuity, it is surprising that we found that the coarse-scale-to-multiscale and fine-scale-to-multiscale models showed similar levels of performance. It may be that there are other reasons as to why evolution has preferred a coarse-scale-to-multiscale progression in biological organisms.

With relatively few exceptions the relationships between development and learning have largely been ignored by the neural computation community. We believe that this is unfortunate. As discussed in the beginning of this chapter, developmental cognitive neuroscientists are in the early phases of formulating a stage-setting theory of neural development. This theory postulates that early events set the stage for later events, meaning that early events are “building blocks” or “stepping stones” for later events. The research project described here highlights the fact that learning and developmental events can be coordinated such that early events provide important functional benefits to later events. In our simulations, we found that orderly changes in scale or resolution in an early computational region, the binocular energy filters, made it easier for a later region, the artificial neural network, to learn to estimate the object and surface disparities present in pairs of visual images. We believe that developmental neurobiologists should not only study individual neural events, but they should also seek to characterize the functional relationships among these events. Computational neuroscientists have an important role to play in this endeavor.

References

- Adelson, E.H. and Bergen, J.R. (1985) Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2, 284-299.
- Atkinson, J. and Braddick, O. (1976) Stereoscopic discrimination in infants. *Perception*, 5, 29-38.
- Barnard, S.T. (1987) Stereo matching by hierarchical, microcanonical annealing (Technical Report 414). Artificial Intelligence Center, SRI International.
- Bishop, C.M. (1995) Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7, 108-116.
- Dominguez, M. and Jacobs, R.A. (2002) Developmental constraints aid the acquisition of binocular disparity sensitivities. Submitted for publication.
- Elman, J.L. (1993) Learning and development in neural networks: The importance of starting small. *Cognition*, 43, 71-99.
- Fleet, D.J., Wagner, H., and Heeger, D.J. (1996) Neural encoding of binocular disparity: Energy models, position shifts, and phase shifts. *Vision Research*, 36, 1839-1857.
- Fox, R., Aslin, R.N., Shea, S.L., and Dumais, S.T. (1980) Stereopsis in human infants. *Science*, 207, 323-324.
- Geman, S., Bienenstock, E., and Doursat, R. (1995) Neural networks and the bias/variance dilemma. *Neural Computation*, 4, 1-58.
- Gray, M.S., Pouget, A., Zemel, R.S., Nowlan, S.J., and Sejnowski, T.J. (1998) Reliable disparity estimation through selective integration. *Visual Neuroscience*, 15, 511-528.
- Greenough, W.T., Black, J.E., and Wallace, C.S. (1987) Experience and brain development. *Child Development*, 58, 539-559.

- Harwerth, R.S., Smith III, E.L., Duncan, G.C., Crawford, M.L.J., and von Noorden, G.K. (1986) Multiple sensitive periods in the development of the primate visual system. *Science*, 232, 235-238.
- Held, R., Birch, E., and Gwiazda, J. (1980) Stereoacuity in human infants. *Proceedings of the National Academy of Sciences USA*, 77, 5572-5574.
- Mallot, H.A., Gillner, S., and Arndt, P.A. (1996) Is correspondence search in human stereo vision a coarse-to-fine process? *Biological Cybernetics*, 74, 95-106.
- Marr, D. and Poggio, T. (1979) A computational theory of human stereo vision. *Proceedings of the Royal Society of London B*, 204, 301-328.
- Matsuoka, K. (1992) Noise injection into inputs in back-propagation learning. *IEEE Transactions on Systems, Man, and Cybernetics*, 22, 436-440.
- McKee, S.P. and Mitchison, G.J. (1988) The role of retinal correspondence in stereoscopic matching. *Vision Research*, 28, 1001-1012.
- Mowforth, P., Mayhew, J.E.W. and Frisby, J.P. (1981) Vergence eye movements made in response to spatial-frequency-filtered random-dot stereograms. *Perception*, 10, 299-304.
- Newport, E.L. (1990) Maturational constraints on language learning. *Cognitive Science*, 14, 11-28.
- Norcia, A. and Tyler, C. (1985) Spatial frequency sweep VEP: Visual acuity during the first year of life. *Vision Research*, 25, 1399-1408.
- Ohzawa, I., DeAngelis, G.C., and Freeman, R.D. (1990) Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science*, 249, 1037-1041.

- Petrig, B., Julesz, B., Kropfl, W., Baumgartner, G., and Anliker, M. (1981) Development of stereopsis and cortical binocularity in human infants: Electrophysiological evidence. *Science*, 213, 1402-1405.
- Piaget, J. (1952) *The Origins of Intelligence in Children*. New York: International Universities Press.
- Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992) *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge, UK: Cambridge University Press.
- Qian, N. (1994) Computing stereo disparity and motion with known binocular cell properties. *Neural Computation*, 6, 390-404.
- Quam, L.H. (1986) Hierarchical warp stereo (Technical Report 402). Artificial Intelligence Center, SRI International.
- Rhode, D.L.T. and Plaut, D.C. (1999) Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, 72, 67-109.
- Shatz, C.J. (1996) Emergence of order in visual system development. *Proceedings of the National Academy of Sciences USA*, 93, 602-608.
- Sietsma, J. and Dow, R.J.F. (1991) Creating artificial neural networks that generalize. *Neural Networks*, 4, 67-79.
- Smallman, H.S. (1995) Fine-to-coarse scale disambiguation in stereopsis. *Vision Research*, 34, 2971-2982.
- Turkewitz, G. and Kenney, P.A. (1982) Limitations on input as a basis for neural organization and perceptual development: A preliminary statement. *Developmental Psychobiology*, 15, 357-368.

Webb, A.R. (1994) Functional approximation by feedforward networks: A least-squares approach to generalization. *IEEE Transactions on Neural Networks*, 5, 363-371.