LETTER Communicated by Richard Zemel

Developmental Constraints Aid the Acquisition of Binocular Disparity Sensitivities

Melissa Dominguez
melissad@cs.rochester.edu
Department of Computer Science, University of Rochester, Rochester, NY 14627, U.S.A.

Robert A. Jacobs
robbie@bcs.rochester.edu
Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, U.S.A.

This article considers the hypothesis that systems learning aspects of visual perception may benefit from the use of suitably designed developmental progressions during training. We report the results of simulations in which four models were trained to detect binocular disparities in pairs of visual images. Three of the models were developmental models in the sense that the nature of their visual input changed during the course of training. These models received a relatively impoverished visual input early in training, and the quality of this input improved as training progressed. One model used a coarse-scale-to-multiscale developmental progression, another used a fine-scale-to-multiscale progression, and the third used a random progression. The final model was nondevelopmental in the sense that the nature of its input remained the same throughout the training period. The simulation results show that the two developmental models whose progressions were organized by spatial frequency content consistently outperformed the nondevelopmental and random developmental models. We speculate that the superior performance of these two models is due to two important features of their developmental progressions: (1) these models were exposed to visual inputs at a single scale early in training, and (2) the spatial scale of their inputs progressed in an orderly fashion from one scale to a neighboring scale during training. Simulation results consistent with these speculations are presented. We conclude that suitably designed developmental sequences can be useful to systems learning to detect binocular disparities. The idea that visual development can aid visual learning is a viable hypothesis in need of study.

1 Introduction

Human infants are born with limited perceptual, motor, and cognitive abilities relative to adults. Within the field of developmental psychology, there

Neural Computation **15, 161–182** (2003)

are at least two perspectives regarding these limitations. The older and more commonplace view is that these limitations are barriers that must be overcome in order for a child to achieve adult function (Piaget, 1952). That is, they are deficiencies or immaturities that serve no positive purpose. A newer view is that these apparent inadequacies are in fact helpful, perhaps necessary, stages in development. Limited mental abilities, according to this theory, reflect simple neural representations that are useful stepping-stones or building blocks for the subsequent development of more complex representations (Turkewitz & Kenney, 1982).

The idea that early developmental stages are useful or necessary precursors to more advanced stages is becoming increasingly studied in the cognitive neurosciences. The development of biological nervous systems is sometimes characterized as using a bootstrapping strategy. Greenough, Black, and Wallace (1987) speculated 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.

Harwerth, Smith, Duncan, Crawford, and von Noorden (1986) provided experimental evidence consistent with this hypothesis by performing 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 earliest stages 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 bootstrapping strategy when attempting to learn a language. Human languages are componential systems in which small linguistic components are systematically combined to form larger linguistic structures. According to Newport's "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 small components that comprise the language. Elman (1993) studied an implementation of this general idea and showed that a recurrent neural network whose memory capacity was initially limited but then gradually increased during the course of training learned aspects of an artificial grammar better than a network whose memory capacity was never limited; that is, the second network's memory capacity was always equal to that of the first network at the end of training. Elman claimed that this outcome supports the idea that starting small 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 article considers the hypothesis that systems learning aspects of visual perception may benefit from the use of suitably designed developmental progressions during training. We report the results of simulations in which four systems were trained to detect binocular disparities in pairs of visual images. Three of the systems were developmental models in the sense that the nature of their input changed during the course of training. These systems received a relatively impoverished visual input early in training, and the quality of this input improved as training progressed. The fourth system was a nondevelopmental model; the nature of 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. The training of the first system, referred to as the coarse-scale-to-multiscale model (model C2M), included a developmental sequence such that the system was exposed only to low-spatial-frequency information at the start of training, and information at higher spatial frequencies was added to its input as training progressed. The training of the second system, referred to as the fine-scale-to-multiscale model (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, and information at lower spatial frequencies was added as training progressed. The third system, referred to as the random-developmental model (model RD), 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. Finally, the fourth system, referred to as the nondevelopmental model (model ND), was not trained using a developmental sequence; it received information at all spatial frequencies throughout the training period.

When comparing the three developmental models with the nondevelopmental model, there are at least two reasonable predictions that one could make about the simulation results. One prediction is that the nondevelopmental model should outperform the developmental models. The nondevelopmental 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 nondevelopmental 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, is that the developmental models would show the better 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. The nondevelopmental model had a greater number of inputs than the developmental models during the early stages of training and, thus, a greater number of modifiable weights. Because learning in neural networks is a search in weight space, the nondevelopmental 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 whose stages are based on spatial frequency content (models C2M and F2M), there are at least three reasonable predictions one could make about the simulation results. One prediction is that model C2M 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 done 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 white dots in the other image. If the images are highly textured, such an approach would be inefficient because there is an intractable number of potential alignments that might need to be checked before the two images are properly aligned (see Figure 1A). If, however, the images are blurred, the fine details that 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 Figure 1B). 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.

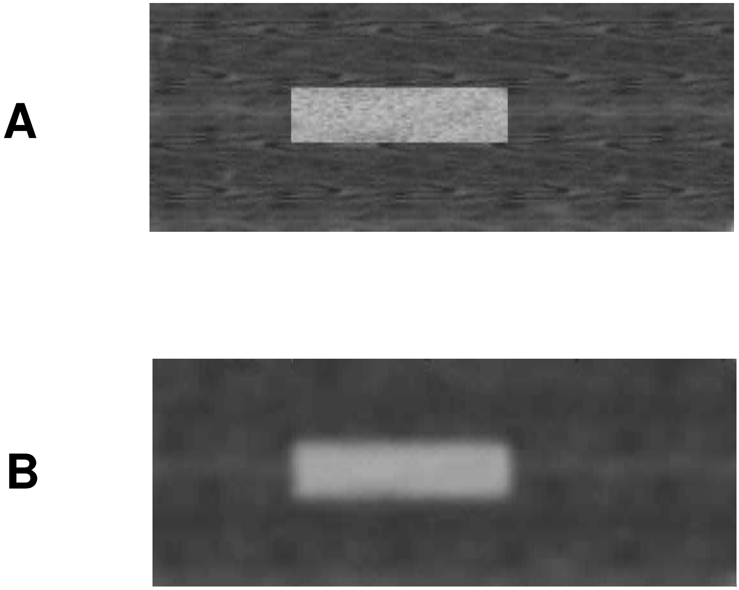

Figure 1: (A) One image from a stereo pair depicting a textured object and background. Due to the fine details in the stereo pair, solving the stereo correspondence problem is relatively difficult. (B) The image from *A* blurred so as to remove the fine details. It is easier to solve the stereo correspondence problem when the stereo pair is blurred because image features tend to be larger, less numerous, and more robust to noise.

This style of processing is commonplace in the computer vision literature. Systems by Marr and Poggio (1979), Quam (1986), and Barnard (1987), among many others, 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 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.

If we assume that human intelligence provides a guide to creating machine intelligence, then a second motivation for the prediction that model C2M 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 discriminate only 1 to 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 eight months of age (Norcia & 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 four months of age (Atkinson & Braddick, 1976; Fox, Aslin, Shea, & Dumais, 1980; Held, Birch, & Gwiazda, 1980; Petrig, Julesz, Kropfl, Baumgartner, & 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 model F2M should perform better than model C2M. 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 bandpass 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, model F2M is initially trained solely with information extracted from high-frequency bandpass 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, 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 counterintuitive 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 & 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 model C2M 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 model F2M to perform best.

Another possible prediction is that models C2M and F2M 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 bandpass 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 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 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 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 (McKee & Mitchison, 1988; Mowforth, Mayhew, & Frisby, 1981; Smallman, 1995). Because human observers do not use a coarse-to-fine strategy or a fine-to-coarse strategy exclusively, we might not expect the exclusive use of a C2M developmental strategy or the exclusive use of a F2M strategy to yield a relative performance advantage.

When comparing the performances of models C2M and F2M with that of model RD, the only reasonable prediction seems to be that model RD should perform no better than the other developmental models, and it is likely that it will perform worse. We have included model RD in this article in order to demonstrate that the use of a developmental sequence does not necessarily lead to performance advantages. Instead, the benefits of a developmental sequence are found only when the sequence incorporates constraints that are useful for learning the desired behavioral task.

This article reports the results of computer simulations comparing the learning performances of the developmental and nondevelopmental models on the task of estimating binocular disparities in novel pairs of images. The results show that the developmental models whose stages are based on spatial frequency content (models C2M and F2M) consistently outperformed the nondevelopmental and random developmental models. We speculate that the superior performance of these models is due to two important features of their developmental progressions: (1) these models were exposed to visual inputs at a single scale early in training, and (2) the spatial scale of their inputs progressed in an orderly fashion from one scale to a neighboring scale. Simulation results consistent with these speculations are presented. We conclude that suitably designed 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 Nondevelopmental Models

Figure 2 illustrates the structure of the developmental and nondevelopmental models. This structure is based on a similar architecture studied by Gray, Pouget, Zemel, Nowlan, & 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 left-most and right-most 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 concerned only with horizontal disparities, as these are the ones that provide information about the three-dimensional configuration of the visual environment.[1] 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 an extension of motion energy filters proposed by Adelson and Bergen (1985). 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:

$$g_L(x, \phi) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \sin(2\pi\omega x + \phi) \tag{2.1}$$

$$g_R(x, \phi) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \sin(2\pi\omega x + \phi + \delta\phi). \tag{2.2}$$

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

---

[1] Vertical disparities provide information about viewing distance and angle of gaze but by themselves do not carry information about the three-dimensional structure of the visual environment. That is, by themselves, they cannot be used for making relative or absolute depth judgments. In contrast, horizontal disparities alone can be used for making relative depth judgments and, when scaled by viewing distance information (perhaps obtained via vertical disparities), for making absolute depth judgments.

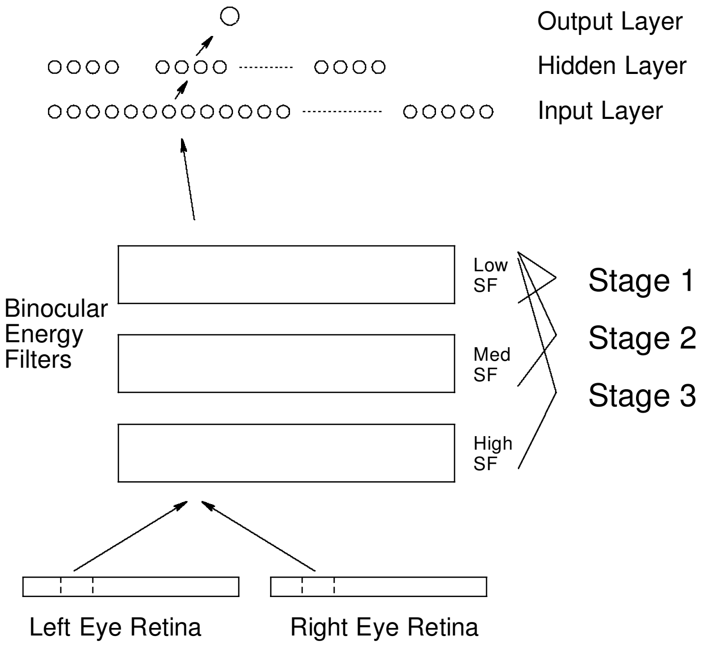

Figure 2: The developmental and nondevelopmental models shared a common structure. The bottom portion of this structure is the left and right retinal images. These images are then filtered by binocular energy filters. The outputs of these filters are the inputs to an artificial neural network that is trained to estimate the disparity present in the images. The model illustrated here is model C2M in which low-spatial-frequency information was received during early stages of training, and information at higher frequencies was added as training progressed.

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, respectively. 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. 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, & 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 that received input from overlapping regions of the retina. At each location, there were 30 complex cells corresponding to three 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 per pixel, 2.0 for 0.125 cycles per pixel, and 4.0 for 0.0625 cycles per pixel. The 10 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,

$$\hat{E}_i(x) = \frac{e^{E_i(x)/\tau}}{\sum_j e^{E_j(x)/\tau}}, \tag{2.3}$$

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. 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 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 disparity present in the right and left retinal images.

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, & 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 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 cover the range of possible binocular disparities uniformly. Training was terminated after 35 iterations through the training set in order to minimize overfitting of the training data. The results reported 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.[2] During the first stage, the neural network portion of the model received only the outputs of complex cells tuned to low spatial frequencies (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; it received the outputs of all complex cells during the third stage. 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. The training of model RD also used a developmental sequence, though this sequence was generated randomly and thus was not based on the spatial frequency tuning of the complex cells. 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 received the outputs of only the complex cells in the first subset. It received the outputs of the cells in the first and second subsets during the second stage of training and the outputs of all complex cells during the third stage. In contrast, the training period for the nondevelopmental model was not divided into separate stages; its neural network received the outputs of all complex cells throughout the training period.

---

[2] The number of iterations in the training stages was roughly optimized using models C2M and F2M and the solid object data set described below. Models with stages of different sizes were tested, and the results suggest that models tended to show worse generalization performance if the number of iterations in the first and second stages was decreased or increased.

3 Data Sets and Simulation Results

The performances of the four models were evaluated on three data sets. These 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. 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 3. Given the right and left images, the task of a model was to estimate the object's disparity.

The upper left graph of Figure 4 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 (the error bars give the standard error of the mean). On average, developmental model C2M had a 16.5% smaller generalization error than the nondevelopmental model (the difference between the mean error rates is statistically significant; $t = 3.77$, $p < 0.002$ using a two-tailed t-test), and a 19.3% smaller error than the random developmental model ($t = 4.60$, $p < 0.001$). Developmental model F2M had a 12.2% smaller error than the nondevelopmental model ($t = 23.74$, $p < 0.001$) and a 15.2% smaller error than the random develop-

Figure 3: Examples of right and left images (top and bottom rows in each panel) from the solid object, noisy object, and planar data sets.

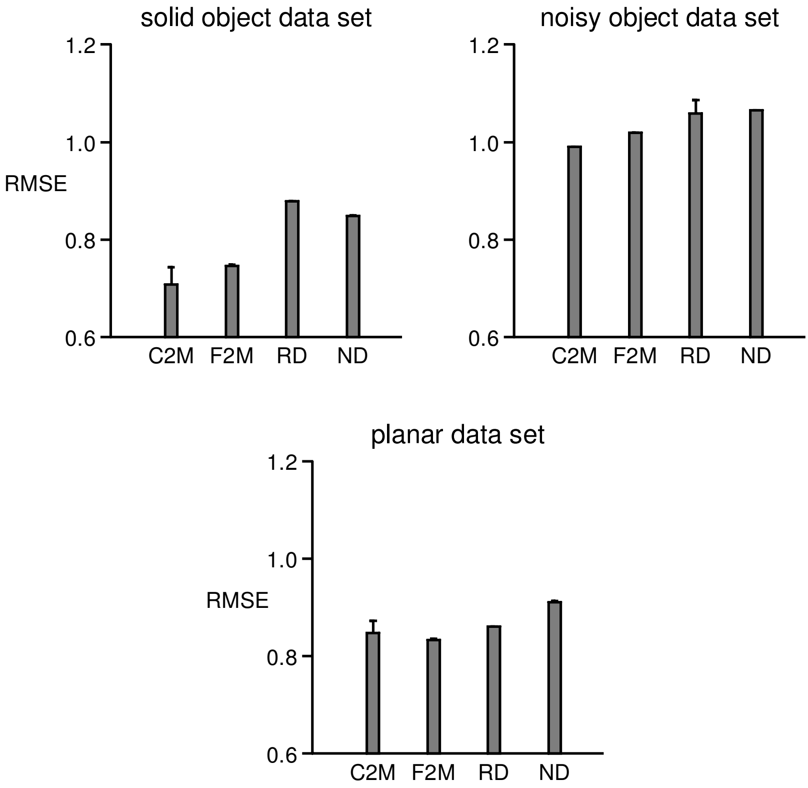

Figure 4: The four models' RMSE on the test set data items after training on the three data sets (the error bars give the standard error of the mean).

mental model ($t = 49.01$, $p < 0.001$). Clearly, the two developmental models whose developmental progressions were organized by spatial frequency content outperformed both the nondevelopmental model and the random developmental model. A statistical comparison between models C2M and F2M shows that their performances were not significantly different.

The variability in model C2M's performance was notably greater than the variability in the performances of the other models. Although it is difficult to know why, it may have to do with the properties of binocular energy units. As discussed by Qian (1994), Zhu and Qian (1996), and Fleet et al. (1996), the responses of these units may peak even when the input disparity is outside the range of disparities to which the unit was thought to be responsive. In such cases, the large response is known as a false peak. False peaks occur for nearly all stimuli, including white noise. Moreover, some of the false peaks will be significantly larger than the peak at the disparity to which the unit was thought to be selective. False peaks tend to occur at integer multiples of the wavelength to which the unit is tuned, meaning that units tuned to low frequencies tend to have false peaks that are relatively far from each other, and units tuned to high frequencies tend to have false peaks near to each other. Consequently, units tuned to low spatial frequencies tend to have false peaks at disparities that are significantly different from the disparity to which they were thought to be responsive, whereas units tuned to high frequencies tend to have false peaks at disparities that are closer to the disparity to which they were thought to be responsive (see Figure 5). The large variability in model C2M's performance may be due to the fact that this model is sometimes misled early in training by false peaks at disparities that are far from the disparities to which units are thought to be selective. This is less likely to be as serious a problem for the other models because they are less likely than model C2M to emphasize information provided by units tuned to low spatial frequencies.

The learning curves for the four models on the solid object data set are illustrated in the upper left graph of Figure 6. The horizontal axis gives the training time in epochs, and the vertical axis gives the RMSE on the test set data items. The solid black line is for model C2M, the solid gray line is for model F2M, the dashed black line is for model RD, and the dotted black line is for model ND. The learning curve for model ND falls quickly within the first few epochs of training and then is relatively flat for the remainder of the training period. In contrast, the developmental models learned relatively slowly, and models C2M and F2M eventually showed the best generalization performance. This result is consistent with the notion described above that apparent inadequacies in performance during early development are not necessarily bad; they sometimes suggest the use of internal representations, which are useful stepping-stones for the subsequent development of advanced behaviors.

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 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 3.

The results are shown in the upper right panel of Figure 4. As before, the developmental models whose developmental progressions were organized by spatial frequency content performed best. On average, model C2M had a 7.1% smaller generalization error than model ND ($t = 448.8$, $p < 0.001$) and a 6.6% smaller error than model RD ($t = 2.46$, $p < 0.05$). Model F2M had a

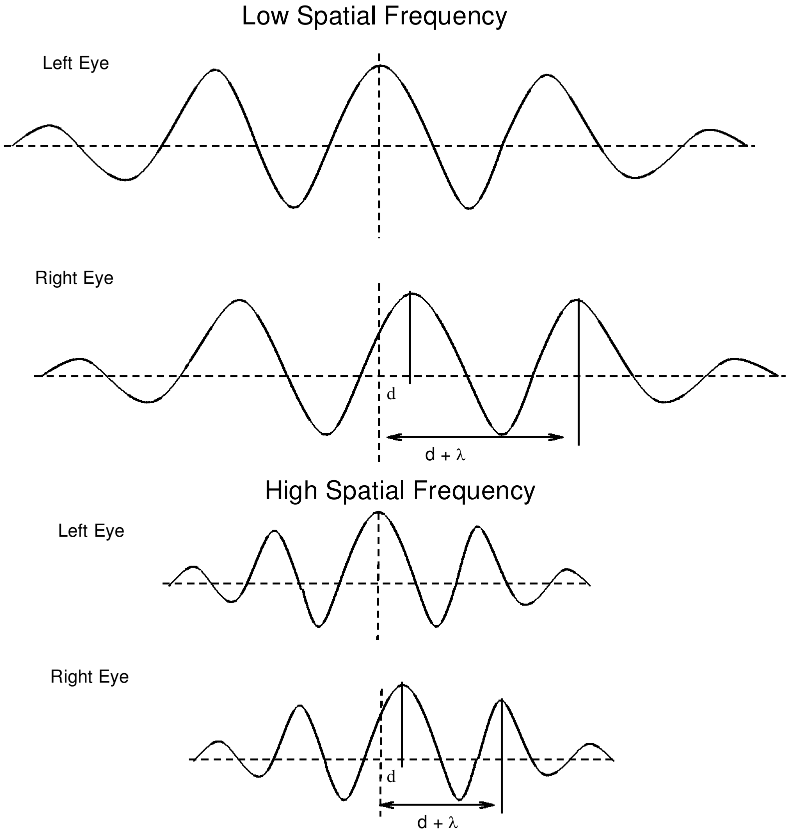

Figure 5: The top two figures illustrate Gabor filters for the left-eye and right-eye images that are tuned to a low spatial frequency (λ is the wavelength). The right-eye Gabor is phase-shifted by *d* units relative to the left-eye Gabor. The sum of the filter responses tends to be large when the disparity in the images is d, $d - \lambda$, or $d + \lambda$. Because the wavelength λ is relatively large, the sum can peak at input disparities (e.g., $d + \lambda$) that are far from the disparity (d) to which the sum was thought to be responsive. As illustrated in the bottom two figures, a similar situation holds for filters tuned to high spatial frequencies, but now the wavelength λ is smaller and the false peaks occur at input disparities that are significantly closer to the disparity to which the sum was thought to be responsive. Consequently, false peaks are more misleading when using filters tuned to low frequencies than when using filters tuned to high frequencies.

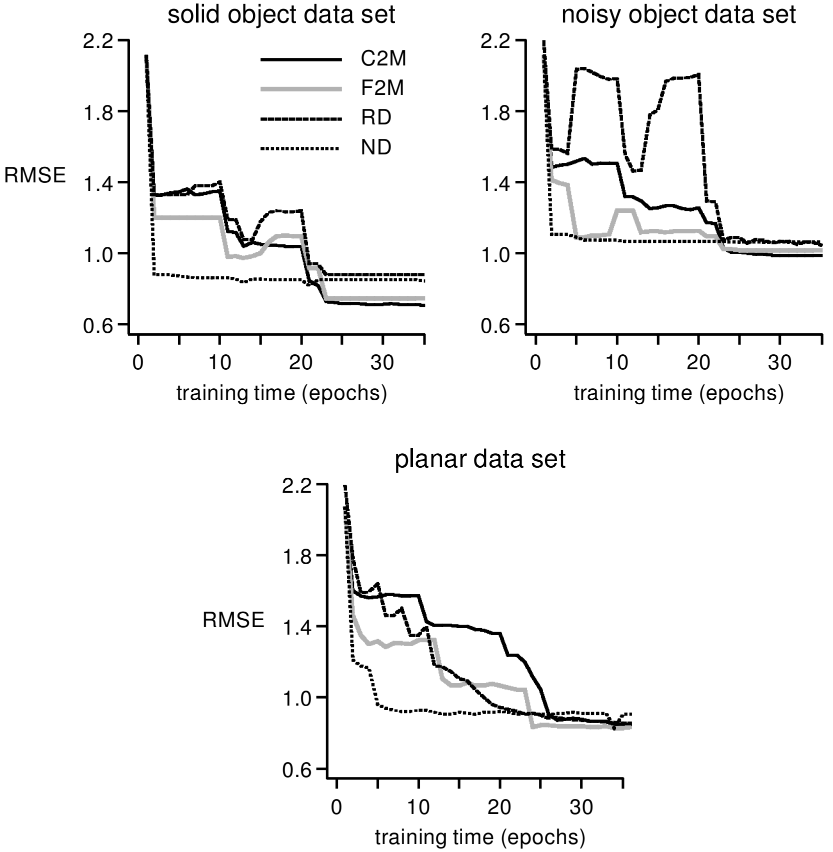

Figure 6: Learning curves for the four models on the three data sets.

4.3% smaller error than model ND ($t = 1566.5, p < 0.001$) and a 3.8% smaller error than model RD (this result is not statistically significant). Comparing the developmental models C2M and F2M, C2M had a 2.85% smaller error ($t = 170.6$, $p < 0.001$). The learning curves for the four models are shown in the upper right graph of Figure 6. A notable feature illustrated by this graph is that the random developmental model had highly unstable levels of performance during the course of training.

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 frontoparallel 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 3.

The results are shown in the bottom graph of Figure 4. Again, the developmental models whose developmental progressions were organized by spatial frequency content tended to outperform the other models. Model C2M had a 6.7% smaller generalization error than model ND ($t = 2.27$, $p < 0.05$) and a 1.3% smaller error than model RD (this result is not statistically significant). Model F2M had an 8.3% smaller error than model ND ($t = 16.84$, $p < 0.001$) and a 3.0% smaller error than model RD ($t = 7.01$, $p < 0.001$). The performances of models C2M and F2M were not statistically different. The bottom graph of Figure 6 gives the learning curves for the four models.

In summary, the simulation results using the three data sets show that the developmental models whose progressions were organized by spatial frequency content (models C2M and F2M) significantly outperformed the nondevelopmental and random-developmental models. To understand the performances of models C2M and F2M better, we analyzed more carefully their behaviors using the solid object data set.

We looked at how these models performed on images with disparities of different sizes (small, midsize, and large disparities) at the end of each developmental stage (stages 1, 2, and 3). These data are given in Figure 7. The three graphs in this figure correspond to the three disparity sizes. The horizontal axis of each graph gives the stage number for each model; the vertical axis gives each model's RMSE at the end of the indicated stage on those test set data items with the indicated disparity size. For purposes of comparison, the graphs also include the data for the nondevelopmental model. Model ND did not significantly change its ability to detect disparities of a given size across training stages; its performance at detecting small, midsize, or large disparities remained nearly constant across the stages. Model C2M improved its ability to estimate disparities of all sizes as training progressed from one stage to the next. In contrast, model F2M appears to have learned about different-sized disparities in each of the different training stages. For example, it learned a great deal about detecting small disparities in stage 1; in stages 2 and 3, it learned more about detecting midsize and large disparities. Not surprisingly, its performance at detecting small disparities degraded as its performance at detecting midsize and large disparities improved.

The question remains as to why models C2M and F2M showed superior performance. One logical possibility is that these models performed well due to the fact that they received relatively few inputs early in training. This possibility can be ruled out, however, because model RD received equally few inputs early in training, but this model did not perform well. We speculate that two important features of the developmental progressions of models C2M and F2M account for their superior performances. First, these models were exposed to visual inputs at a single scale early in training. Model C2M received only coarse-scale information at the start of training; model F2M received only fine-scale information. In contrast, models RD and ND received information at all spatial scales at all stages of training.

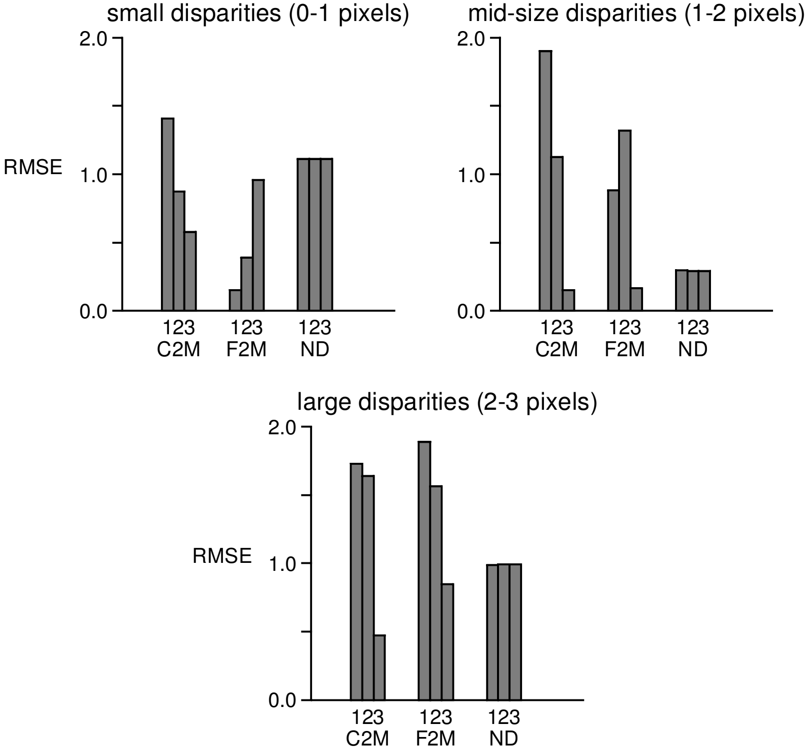

Figure 7: Performance of models C2M, F2M, and ND on images with disparities of different sizes (small, midsize, and large disparities) at the end of each developmental stage (stages 1, 2, and 3). Training and test data items came from the solid object data set.

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

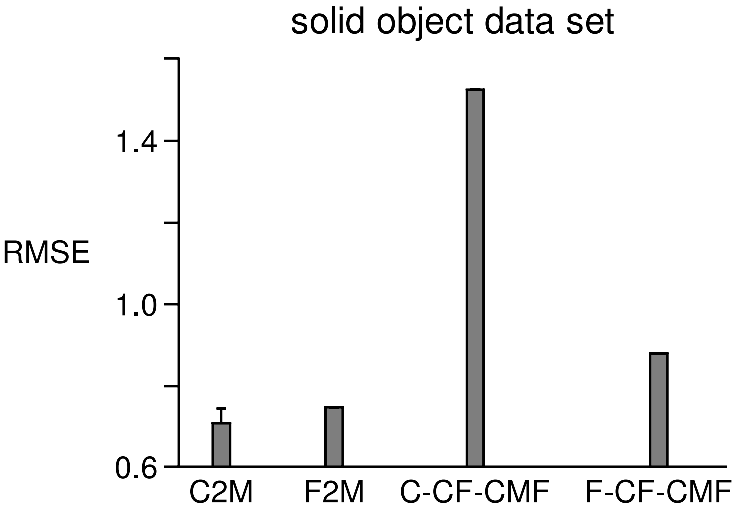

Figure 8: RMSE of models C2M, F2M, C-CF-CMF, and F-CF-CMF on the test items from the solid object data set.

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 early in training. In stage 2, it received the outputs of complex cells tuned to low and high frequencies, and it received the outputs of cells tuned to low, medium, and high frequencies in stage 3. 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 2, and the outputs of all complex cells in stage 3.

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. These data are consistent with the conjecture that it is advantageous to a developmental system for the spatial scale of its inputs to progress in an orderly fashion from one scale to a neighboring scale.

4 Summary and Conclusion

This article has considered the hypothesis that systems learning aspects of visual perception may benefit from the use of suitably designed developmental progressions during training. We reported the results of simulations in which different systems were trained to detect binocular disparities in pairs of visual images. Three of the systems were developmental models in the sense that the nature of their input changed during the course of training. These systems received a relatively impoverished visual input early in training, and the quality of this input improved as training progressed. The fourth system was a nondevelopmental model; the nature of its input remained constant during the course of training. The results show that the developmental models whose stages were based on spatial frequency content, models C2M and F2M, consistently outperformed the nondevelopmental and random developmental models. We speculate that the superior performance of these models is due to two important features of their developmental progressions: (1) these models were exposed to visual inputs at a single scale early in training, and (2) the spatial scale of their inputs progressed in an orderly fashion from one scale to a neighboring scale. Simulation results consistent with these speculations were presented. We conclude that suitably designed developmental sequences can be useful to systems learning to detect binocular disparities, in accord with the less-is-more view of development. Moreover, the idea that visual development can aid visual learning is a viable hypothesis in need of study.

With relatively few exceptions, the relationship between development and learning has been ignored by the neural computation community. We believe that this is unfortunate. It is well known that systems learn best when they are suitably constrained through the use of domain knowledge. Learning systems are inherently faced with the bias-variance dilemma (Geman, Bienenstock, & Doursat, 1995). Systems with little or no bias are often capable of learning many different sets of training items. Unfortunately, they tend to interpolate in unpredictable ways and thus generalize poorly to novel data items. In contrast, systems that are constrained through the use of domain knowledge and thus have large bias are not able to learn as wide a variety of training sets. However, they tend to show better generalization performance and less variable generalization performance when exposed to those training sets that they can adequately learn. The design of appropriate developmental progressions through the use of domain knowledge provides researchers with an effective means of biasing their learning systems so as to enhance their performances.

Acknowledgments

This work was supported by NSF graduate fellowship DGE9616170 and by NIH research grant R01-EY13149.

References

Adelson, E. H., & 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., & Braddick, O. (1976). Stereoscopic discrimination in infants. *Perception, 5*, 29–38.

Barnard, S. T. (1987). *Stereo matching by hierarchical, microcanonical annealing* (Tech. Rep. 414). Menlo Park, CA: Artificial Intelligence Center, SRI International.

Bishop, C. M. (1995). Training with noise is equivalent to Tikhonov regularization. *Neural Computation, 7*, 108–116.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 43*, 71–99.

Fleet, D. J., Wagner, H., & 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., & Dumais, S. T. (1980). Stereopsis in human infants. *Science, 207*, 323–324.

Geman, S., Bienenstock, E., & 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., & Sejnowski, T. J. (1998). Reliable disparity estimation through selective integration. *Visual Neuroscience, 15*, 511–528.

Greenough, W. T., Black, J. E., & Wallace, C. S. (1987). Experience and brain development. *Child Development, 58*, 539–559.

Harwerth, R. S., Smith, E. L., Duncan, G. C., Crawford, M. L. J., & von Noorden, G. K. (1986). Multiple sensitive periods in the development of the primate visual system. *Science, 232*, 235–238.

Held, R., Birch, E., & Gwiazda, J. (1980). Stereoacuity in human infants. *Proceedings of the National Academy of Sciences USA, 77*, 5572–5574.

Mallot, H. A., Gillner, S., & Arndt, P. A. (1996). Is correspondence search in human stereo vision a coarse-to-fine process? *Biological Cybernetics, 74*, 95–106.

Marr, D., & 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., & Mitchison, G. J. (1988). The role of retinal correspondence in stereoscopic matching. *Vision Research, 28*, 1001–1012.

Mowforth, P., Mayhew, J. E. W., & 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., & 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., & 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., & 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., & Flannery, B. P. (1992). *Numerical recipes in C: The art of scientific computing*. Cambridge: 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* (Tech. Rep. 402). Menlo Park, CA: Artificial Intelligence Center, SRI International.

Rohde, D. L. T., & 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., & 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., & 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 feed-forward networks: A least-squares approach to generalization. *IEEE Transactions on Neural Networks, 5*, 363–371.

Zhu, Y., & Qian, N. (1996). Binocular receptive field profiles, disparity tuning, and characteristic disparity. *Neural Computation, 8*, 1611–1641.

Received November 14, 2001; accepted June 4, 2002.