

Visual Development and the Acquisition of Binocular Disparity Sensitivities

Melissa Dominguez

MELISSAD@CS.ROCHESTER.EDU

Department of Computer Science, University of Rochester, Rochester, NY 14627 USA

Robert A. Jacobs

ROBBIE@BCS.ROCHESTER.EDU

Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627 USA

Abstract

This paper 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 three different artificial neural network models were trained to detect binocular disparities in pairs of visual images. Two of the models were developmental models in the sense that the nature of their training input changed during the course of training (either a coarse-scale-to-multiscale progression or a fine-scale-to-multiscale progression). The third model was a non-developmental model in the sense that its training input remained constant during the training period. The simulation results show that the two developmental models consistently outperformed the non-developmental model. We conclude that developmental sequences during training can be useful to systems learning to detect binocular disparities. The idea that developmental progressions can aid visual learning is a viable hypothesis in need of future study.

1. Introduction

Human infants are born with limited perceptual, motor, and cognitive abilities relative to adults. There are at least two perspectives within the field of developmental psychology regarding these limitations. One view is that these limitations are barriers which must be overcome in order for a child to achieve adult function (Piaget, 1952). According to this view, men-

tal limitations are immaturities or deficiencies which serve no positive purpose. An alternative view which is recently gaining in popularity is that these apparent inadequacies are in fact helpful, perhaps necessary, stages in development. 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).

For example, a bootstrapping strategy has been hypothesized to be used by children learning 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. Elman (1993) studied an implementation of this general idea. 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. According to Elman, 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 and so it is difficult to know how to interpret these findings.

This paper considers the hypothesis that systems learning aspects of visual perception may benefit from the use of suitably designed developmental progres-

sions during training. We report the results of simulations in which three different systems were trained to detect binocular disparities in pairs of visual images. Two of the systems are developmental models in the sense that the nature of their training input changed during the course of training, whereas the third system is a non-developmental model in the sense that its training input remained constant during the training period. 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* (or 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 the input as training progressed. The training of the second system, referred to as the *fine-scale-to-multiscale model* (or model F2M), also included a developmental sequence. This system received high spatial frequency information at the start of training; information at lower spatial frequencies 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.

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 two 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. The intuitive assumption that more information is better than less information leads to the prediction that the non-developmental model ought to perform best. An alternative prediction 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 knowledge representations, then the developmental models ought to have an advantage.

The simulation results show that the two developmental models, models C2M and F2M, consistently outperformed the non-developmental model on the task of estimating binocular disparities in novel pairs of images. In addition, because model C2M performed as well as or better than model F2M, the results indicate that a coarse-scale-to-multiscale developmental progression may be preferable to a fine-scale-to-multiscale

progression. 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 developmental progressions can aid visual learning is a viable hypothesis in need of future study.

There are many possible developmental systems for the task of binocular disparity detection that one might consider. We have focused on the coarse-scale-to-multiscale model and the fine-scale-to-multiscale model for the following reasons. A motivation for the coarse-scale-to-multiscale developmental sequence is the fact that human infants show a related developmental progression. Human 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. Adults with normal vision (so-called 20/20 vision) can discriminate approximately 30 cycles per degree of arc. Newborns, however, 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 (Fox, Aslin, Shea, and Dumais, 1980; Held, Birch, and Gwiazda, 1980). 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.

A second motivation for our study of a coarse-scale-to-multiscale developmental sequence comes from the field of computer vision where a coarse-to-fine processing strategy is frequently used. Systems by Marr and Poggio (1979), Quam (1986), and Barnard (1987), among many others, initially search for stereo 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 on the basis of information from one or more higher-resolution pairs of images. There is, however, an important difference between the use of a coarse-to-fine strategy by computer vision researchers and our use of the closely related coarse-scale-to-multiscale strategy. Computer vision researchers use a coarse-to-fine sequence when processing each individual pair of images. In contrast, we used

the coarse-scale-to-multiscale sequence while training a learning system that was exposed to many pairs of images. Early in training the system was only provided with low-resolution pairs of images; information at progressively higher resolutions was added at subsequent stages of training. 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.

Although computer vision researchers frequently use a coarse-to-fine processing strategy, psychologists have discovered that human observers often do not. Malot, 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 (e.g., McKee and Mitchison, 1988; Mowforth, Mayhew, and Frisby, 1981; Smallman, 1995). These psychophysical experiments provide a motivation for studying model F2M which was exposed only to fine-scale information during early stages of training, with coarser scale information added to its input as training progressed.

Section 2 of this paper describes the binocular energy filters applied to the left and right retinal images, and describes the structure and training of the three models. Section 3 compares the performances of the models on three data sets. A summary and conclusions are provided in Section 4.

2. Developmental and Non-Developmental Models

The structure of the developmental and non-developmental models is based on a similar architecture studied by Gray, Pouget, Zemel, Nowlan, and Sejnowski (1998). The retinal layer of each model consisted of two one-dimensional arrays 62 pixels in length for the left and right eye images (see Figure 1). Each retina was treated as if it were shaped like a circle; in order to avoid edge artifacts the leftmost and rightmost pixels were regarded as neighbors. Although one-dimensional retinas are a simplification, their use is justified by the fact that the model was concerned

only 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.

Ohzawa, DeAngelis, and Freeman (1990) proposed binocular energy filters as a way of modeling the binocular sensitivities of simple and complex neurons 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:

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

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

Each function is a sinusoid multiplied by a Gaussian envelope, where 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 simple cell's output 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 with the same phase offsets, though with different base phases. Because the base phases of these simple cells form quadrature pairs (the base phases are $0, \pi/2, \pi,$ and $3\pi/2$), the complex cell's output is relatively insensitive to the exact position of a disparity within its receptive field.

The models that we simulated had 35 receptive-field locations which received input from overlapping regions of the retina. There were 30 complex cells at each of these locations corresponding to 3 spatial frequencies and 10 phase offsets at each frequency. The three spatial frequencies were each separated by an

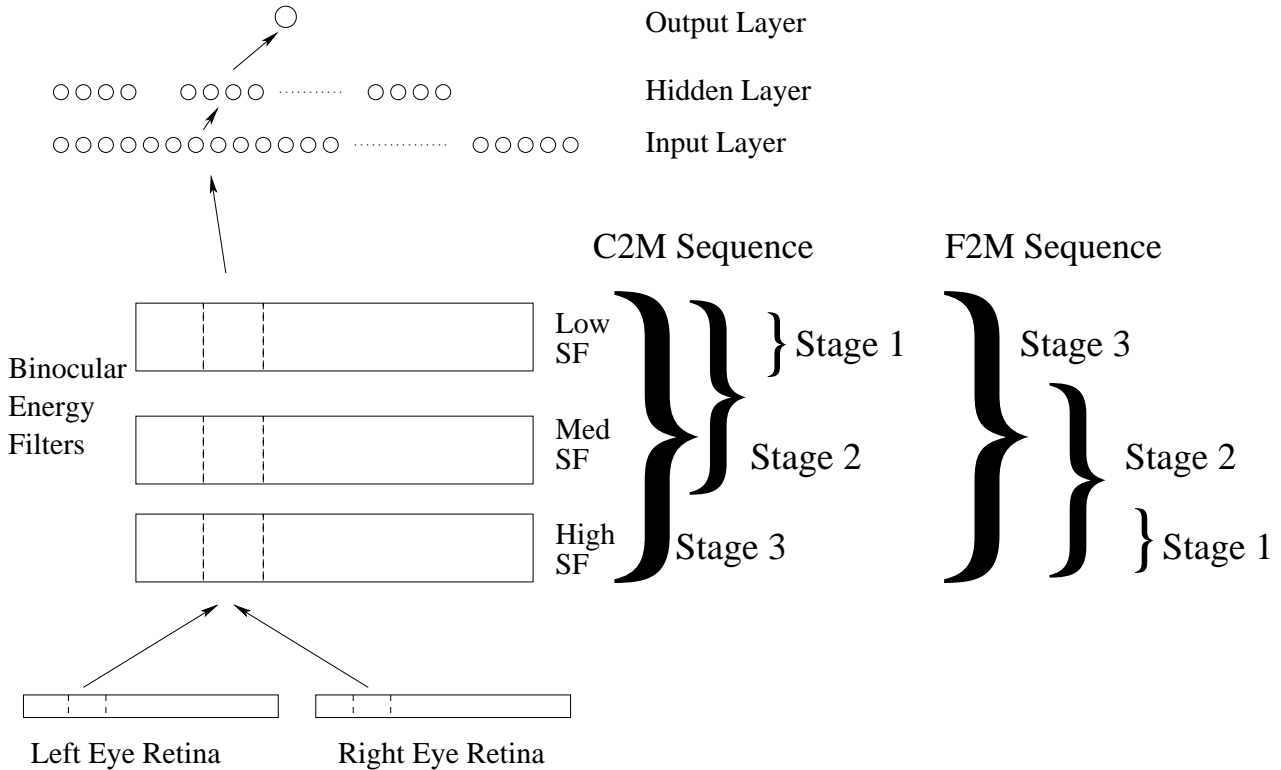


Figure 1. The developmental and non-developmental models shared a common structure. Illustrated here is the coarse-scale-to-multiscale model in which low spatial frequency information was received during early stages of training and information at higher frequencies was added as training progressed, and the fine-scale-to-multiscale model in which high spatial frequency information was received during early stages of training and information at lower frequencies was added as training progressed. In the non-developmental model, all spatial frequencies were present throughout training.

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:

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

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, thus the filtering described above acted as a preprocessing step. 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 disparity present in the right and left 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, 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 in-

coming 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 dramatically decreased the number of modifiable weight values in the network. It also 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 disparities. Training was terminated after 35 iterations through the training set in order to prevent 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 received only the outputs of complex cells tuned to low spatial frequencies (the outputs of the other complex cells were set to zero). The network received the outputs of complex cells tuned to low and medium spatial frequencies during the second stage; 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. Its network received the outputs of complex cells tuned to high spatial frequencies during the first stage. 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 model ND 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.

Images in the *solid object data set* 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 2. Given the right and left images, the task of a model was to estimate the object's disparity.

The results are shown in the leftmost graph of Figure 3. 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 an 12.2% smaller error than the non-developmental model ($t = 23.74$, $p < 0.001$). Clearly, the two developmental models outperformed the nondevelopmental model. A statistical comparison between the developmental models C2M and F2M shows that their performances were not significantly different.

The leftmost graph of Figure 4 gives the learning curves for the three models. The horizontal axis gives the training time in epochs, or iterations through the training set, and the vertical axis gives the RMSE on the test set data items. The solid line is for model C2M, the dashed line is for model F2M, and the dotted line is for model ND. Interestingly, the two developmental models learned slowest but eventually showed the best generalization performance. We believe that this result is consistent with the notion described above that apparent inadequacies in performance during early development are not necessarily bad; they can sometimes suggest the use of knowledge representations which are useful stepping-stones for the subsequent development of advanced behaviors.

Images in the second data set, referred to as the

Solid Object

Noisy Object

Planar

Figure 2. Examples of right and left images (top and bottom rows in each panel) from the solid object, noisy object, and planar data sets.

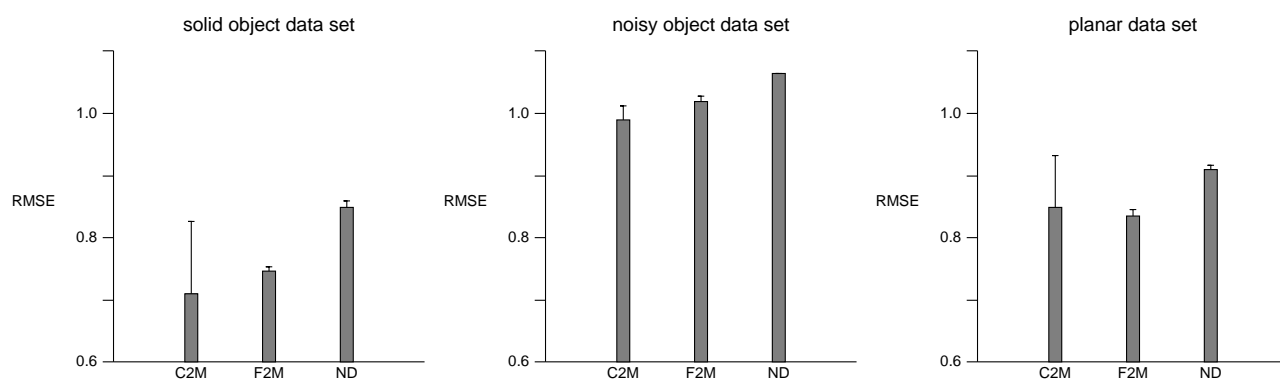


Figure 3. The three models' root mean squared errors (RMSE) on the test set data items after training on the three data sets (the error bars give the standard deviations).

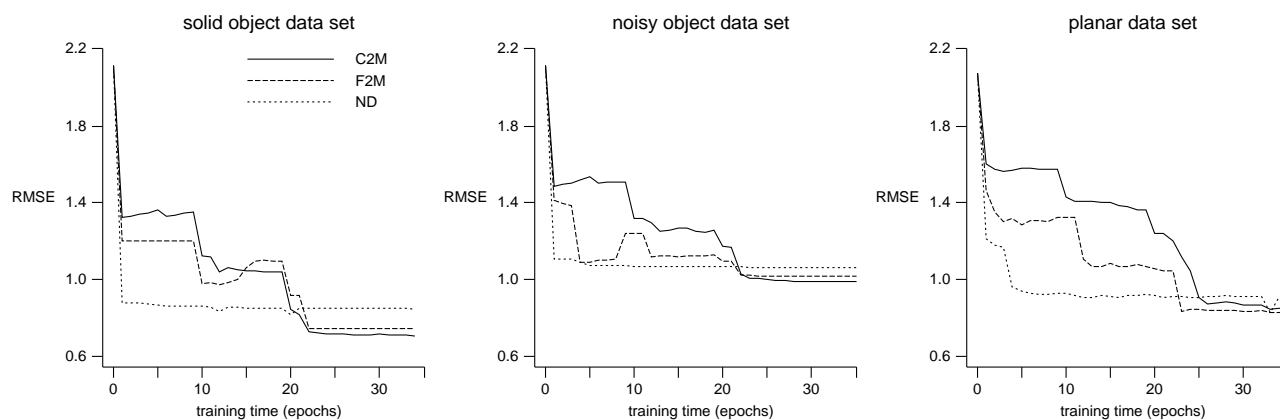


Figure 4. The three models' root mean squared errors (RMSE) on the test set data items at the end of each iteration through the training set when trained on the three data sets.

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 right and left images to an estimate of the object's disparity. An example of a right and left image is shown in the middle panel of Figure 2.

The results are shown in the middle panel of Figure 3. In short, 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 learning curves for the three models are shown in the middle graph of Figure 4.

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 right and left images, the task was to estimate the shift. An example of a right and left image is shown in the bottom panel of Figure 2.

The rightmost graph of Figure 3 gives the results. 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. The rightmost graph of Figure 4 gives the learning curves for the three models.

Overall, the simulation results clearly reveal that the developmental models performed significantly better than the non-developmental model on the data sets evaluated here. In addition, the results suggest that model C2M is preferable to model F2M; the coarse-scale-to-multiscale developmental sequence of model

C2M yielded performance that was as good as or better than that of the fine-scale-to-multiscale sequence of model F2M. These results are consistent with the view that apparent inadequacies in performance during early development are not necessarily bad; they can sometimes suggest the use of knowledge representations which are useful stepping-stones for the subsequent development of advanced behaviors.

4. Conclusions

With relatively few exceptions the relationship between development and learning has largely been ignored by the machine learning community. We believe that this is unfortunate. The simulation results reported here show that suitably designed developmental sequences can be useful to systems learning to detect binocular disparities. Moreover, these results suggest that the idea that developmental progressions can aid visual learning (and perhaps other forms of learning as well) is a viable hypothesis in need of future study.

It is well known in the machine learning literature 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, and Doursat, 1995). Systems with little or no bias typically have highly variable generalization performance. Because they are relatively unconstrained, they are 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 with large bias frequently show less variability in their generalization performance. These systems are constrained through the use of domain knowledge and, thus, are not able to learn as wide a variety of training sets. However, they tend to show better 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 machine learning researchers with an effective means of enhancing the learning performances of their systems.

Acknowledgments

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

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.
- Barnard, S.T. (1987) Stereo matching by hierarchical, microcanonical annealing (Technical Report 414). Artificial Intelligence Center, SRI International.
- 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.
- 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.
- 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) Maturation 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.
- 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.
- Rohde, 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.
- 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.