

Interactions Between Development and Learning During the Acquisition of Binocular Disparity Sensitivities

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

Abstract

This paper investigates the use of developmental progressions in the acquisition of binocular disparity sensitivity. In an earlier paper we presented results of simulations comparing a non-developmental neural network model and two developmental neural network models trained to detect binocular disparities and concluded that the developmental models consistently performed better[1]. Here we report the results of simulations on additional neural network models, compare them to the three original models and perform additional analysis. We conclude that the benefits of development are not solely due to limiting the size of the input in early stages, and that in order for development to be helpful to the task of binocular disparity detection, the developmental progressions should be designed to take advantage of known simplifications.

1 Introduction

Human infants are born with immaturities in their perceptual, motor, and cognitive systems relative to adults that effectively act as limitations to those abilities. In the developmental psychology literature there are at least two schools of thought about these limitations. The more traditional view, as espoused by Piaget[2] and others, is that these immaturities are obstacles that must be overcome in order for the infant to become an adult. That is, the initial limitations are in effect a handicap which serve no positive purpose. Turkewitz and Kenney [3] proposed an alternate perspective on these initial limitations. According to this increasingly popular view, what appear to be inadequacies are in fact advantageous. They effectively simplify the world allowing for the building of simple neural representations that act as “building blocks” or “stepping stones” for the later development of more complex neural representations.

Such bootstrapping strategies have been hypothesized in children’s language acquisition. Newport’s[4]

“Less is More” hypothesis states that children’s limited memorial and attentional abilities enable children to segment language into its smaller componential parts, and that adults’ difficulty in learning new language stems in part from a poor ability to do this segmentation [5]. Elman [6] built a system based on this general idea. He compared the performance of a recurrent neural network with initially limited memory capacity that was increased over training, to a network whose memory was always equal to that of the first network at the end of training. He showed that the first network learned aspects of a grammar better than the second, thus supporting the idea that “starting small” is important to language learning. Rhode and Plaut [7] were unable to reproduce these results, however, so it is difficult to know how to interpret them.

This paper considers the hypothesis that systems learning aspects of visual perception may benefit from suitably designed developmental sequences. In earlier research[1] we performed simulations on three neural network models: model C2M(Coarse-to-Multiscale), model F2M(Fine-to-Multiscale) and model ND(Non-Developmental). We compared the performance of these three models on acquisition of binocular disparity sensitivities. We concluded that the two developmental models outperformed the non-developmental model in all of our simulations, and thus that the developmental approach was advantageous for problems of this nature. But the question remained, was the performance improvement observed merely due to early limitations on input size, or was it due to the principled design of models C2M and F2M? We speculate that the important features of the developmental sequences used in C2M and F2M are that (1) the visual inputs to which these models are exposed are of a single spatial frequency scale early in training, and (2) the scale of the inputs progresses to neighboring scales

in an orderly fashion between stages of training. This paper presents simulation results consistent with this hypothesis and provides further analysis of the results of the original paper.

2 Previous Work

In the original paper, we compared three models: a non-developmental model (ND), a Coarse-to-Multiscale model(C2M), and a Fine-to-Multiscale model(F2M). Models C2M and F2M were developmental in that the nature of their input changed during the course of training. The input to model ND did not change during training. The developmental models had inspiration both from existing computer vision approaches to disparity detection and from human psychophysical studies.

Model C2M first received only coarse grained, or low spatial frequency, data. At subsequent stages of training, data at increasingly fine scales, or higher spatial frequencies, was added to its input. This was intended to model a similar progression in human infants. At birth, infant spatial frequency acuity is roughly 1/15 - 1/30 of that of an adult with normal eyesight. They are sensitive only to low spatial frequencies – that is, they can not see fine details. Acuity improves approximately linearly from these low levels at birth to near adult levels by about 8 months of age[8]. Interestingly, infants are acquiring other visual skills during this time period; in particular, sensitivity to disparities appears around 4 months of age [9, 10]. We speculate that this early poor acuity aids in the acquisition of disparity detection in infants, and thus could also be helpful to an artificial neural network learning the same task.

The field of computer vision, where coarse-to-fine processing strategies are often used, provides further motivation for the C2M model. Systems using this strategy (see [11, 12, 13] and many others) typically search for stereo correspondences first in a pair of low resolution images. Low resolution images have fewer, larger, more robust features, and thus the search is constrained and allows for a rough estimate to be made. This estimate is then refined using one or more higher resolution image pairs.

The F2M model received only fine-scale, or high spatial frequency, information early in training, and progressively added coarser, lower frequency, data as training proceeded. This model was inspired by adult human psychophysical studies[14] that show that humans can use high spatial frequency information to disambiguate low spatial frequency information, and further that we are not always able to do the reverse. This would indicate that humans are, at least on oc-

casional, using a fine-to-coarse processing strategy (see also [15, 16, 17]). This is also similar to some computer vision systems which use edge-detection(a form of high-frequency band-pass filtering) as an initial step in solving the stereo correspondence problem. Again this is because edges are often large, sparse, and robust to noise.

Model ND was intended as a control. In contrast to the developmental models, its input did not change during the course of training. It received all inputs from the beginning.

2.1 The Models

For a considerably more in-depth discussion of the models please see [1]. All three models shared a common structure, as shown in Figure 1. This structure is based on a similar architecture studied by Gray, Pouget, Zemel, Nowlan, and Sejnowski [18]. The retinal layer consisted of two one-dimensional arrays of 62 pixels for the left and right images.

The next layer consisted of binocular energy filters. As proposed by Ohzawa, DeAngelis, and Freeman [19], binocular energy filters model the binocular sensitivities of simple and complex cells in the primary visual cortex in primates. In these filters, a simple cell receives input from subunits 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 Gaussian windowed sinusoid, 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 the left subunit is offset from the phase of the right by $\delta\phi$. The outputs of these subunits are summed, then this sum is half-wave rectified and squared to produce the output of the simple cell. The output of a complex cell is the sum of the outputs of four simple cells with the same phase offsets, and with base phases that form quadrature pairs. This is the final output of the binocular energy filter.

In our simulations, there were 35 receptive field locations which received data from overlapping sections of the retinae. At each receptive field location there were 30 binocular energy filter units corresponding to three spatial frequencies(low, medium and high) and 10 phase offsets for each frequency. The outputs from

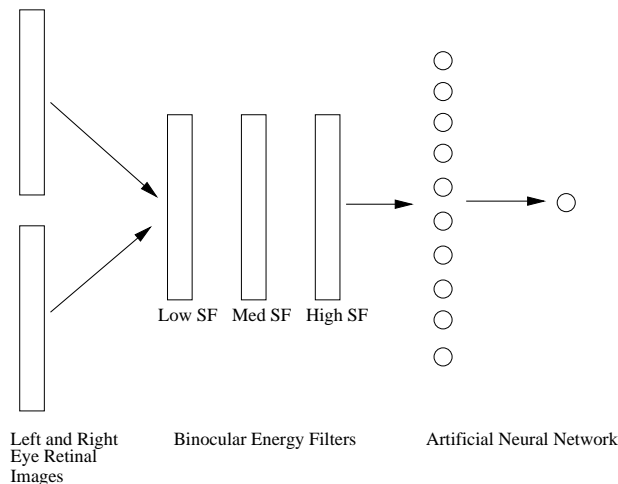


Figure 1: The developmental and non-developmental models shared a common structure. Information from the left and right eye retinæ was processed through binocular energy filters tuned to low, medium and high spatial frequencies before being passed to the neural network.

the binocular energy filter layer were inputs to an artificial neural network (see Figure 2).

For the developmental models, the training period was divided into three stages, with the training data qualitatively different for each stage. For model C2M, the training proceeded in a coarse-scale-to-multiscale progression. During the first stage, the neural network portion of the model received only the outputs of the binocular energy filters tuned to low spatial frequencies. In the second stage it received the outputs of the filters tuned to low and medium spatial frequencies, and in the final stage it received all outputs. Model F2M was trained in a fine-scale-to-multiscale progression. In the first stage it received the outputs of only the filters tuned to high spatial frequencies, in the second stage it received the outputs of the high and medium frequency filters and in the final stage it received all outputs. The non-developmental model (ND) received all outputs throughout training.

2.2 Data Sets and Simulation Results

The three models were evaluated by their performance on three data sets (based on data sets used by Gray et al. [18]). All data sets were gray scale images with luminance values between 0 and 1 and disparities between 0 and 3 pixels. Results can be seen in Figure 4 and will be discussed in later sections.

The images in the *solid object data set* consisted of a single light (luminance between 0.0 and 0.1) or dark (0.9 and 1.0) object on a gray background (0.5). The object’s size, location, and disparity were randomly chosen real numbers. When object edges fell between pixel locations, the gray-scale values were in-

terpolated. An example of right and left images is given in the top panel of Figure 3. Given right and left images, the task was to estimate the disparity between the object’s location in the two images.

In the *noisy object data set* images again consisted of a single object on a background. However in this data set both the background and the object pixels were randomly generated. The object’s size, location, and disparity were again randomly generated. The background had zero disparity. An example of right and left images is given in the middle panel of Figure 3. Given right and left images, the task was to estimate the disparity between the object’s location in the two images.

The *planar data set* consisted of binary images of a fronto-parallel plane with a uniform disparity. An example of right and left images is given in the bottom panel of Figure 3. Given right and left images, the task was to estimate the overall disparity between the two images.

Overall, the two developmental models did significantly better (according to a two-tailed t-test, $p < 0.002$ in all cases) than the non-developmental in each data set. These experiments demonstrated that developmental sequences can be beneficial to the process of acquiring binocular disparity sensitivities, but the question remained of whether these benefits were due to the careful design of those sequences or merely to the early limitation on the *size* of the training data items.

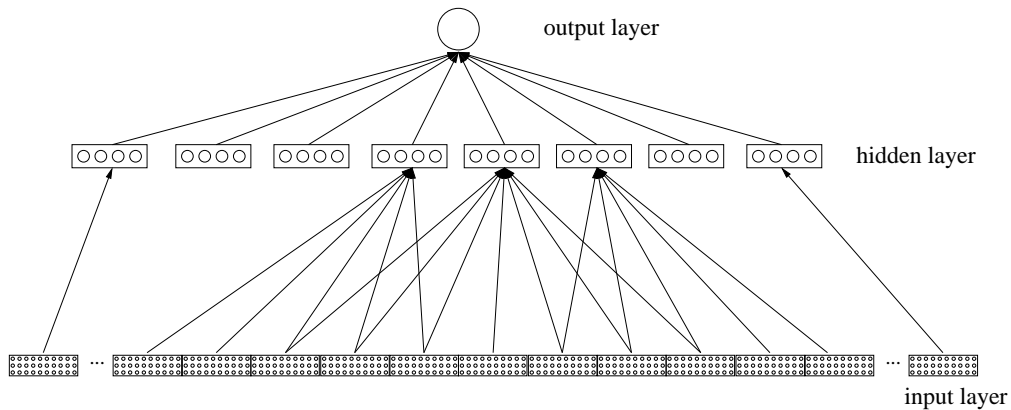


Figure 2: The structure of the artificial neural network portion of the developmental and non-developmental models. The activations of the input units were set to the normalized outputs of the complex cells. The 32 hidden units were organized into 8 groups with 4 units each. Weight sharing among the hidden units' incoming and outgoing weights provided the network with a degree of translation invariance. The output unit was trained to estimate the disparity present in the left and right images.

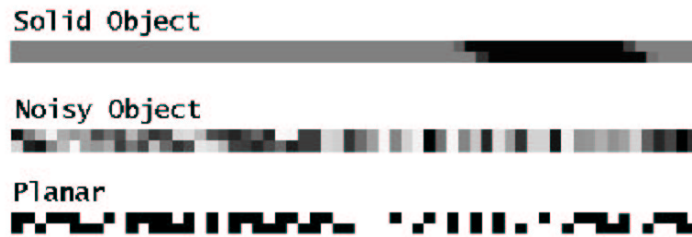


Figure 3: Examples of left and right eye images (top and bottom rows) from the three data sets. Top panel: in going from the top to the bottom images, the dark object is shifted to the right. Middle panel: the noisy object is towards the left edge of the images. In going from the top to the bottom images the noisy object is shifted by two pixels to the right. Bottom panel: the entire top images is shifted by two pixels to the right in order to form the bottom image.

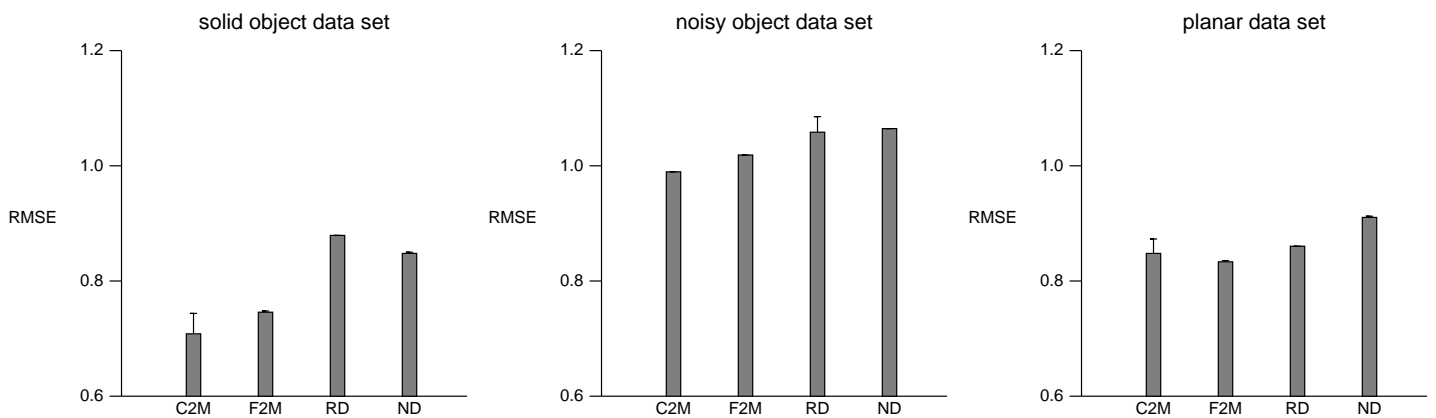


Figure 4: The three models' root mean squared errors (RMSE) on the test set data items after training on the three data sets (the error bars give the standard deviations).

3 Additional Developmental Models

To determine whether or not the benefits shown in the first paper could be explained away by mere limitations on input size, we designed a third developmental model. Model RD (randomized development) received the outputs from a semi-random third of the disparity energy filters at the first stage of training and had an additional semi-random third added at each subsequent stage. The randomness was constrained such that each third contained exactly one energy filter output from each phase difference and receptive field location. The results can be seen in Figure 4. Model RD showed none of the benefits of development that models C2M and F2M displayed, performing nearly as poorly or even worse than model ND. Thus clearly the limitations on size alone are not enough to ensure profit from developmental sequences.

We speculated that the benefits witnessed in models C2M and F2M were due to two factors. Namely (1) each training stage consisted of outputs from a single spatial frequency, and (2) training proceeded in an orderly fashion from scale to neighboring scale. Condition (1) would allow 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. If condition (2) is satisfied, when the system receives inputs at a new spatial scale, it is close to a scale with which the system is already familiar.

To test the importance of condition (2) two additional models were constructed and tested on the solid object data set. These two models satisfied condition (1) but not (2); that is, each stage consisted of outputs from disparity energy filters at a single spatial frequency, but the progression from stage to stage was out of order with respect to scale. The two models were model C-CF-CMF (coarse-coarse and fine-coarse, medium, and fine) which received first low spatial frequency information, then low and high, and finally all three; and model F-CF-CMF (fine-coarse and fine-coarse, medium, and fine) which received first high spatial frequency information, then low and high, and finally all three. Since we conjecture that condition (2) is, indeed, important, we would predict that these two models would perform poorly.

The results can be seen in Figure 5 which shows the performance of models C2M, F2M, C-CF-CMF, and F-CF-CMF at the end of training on the solid object data set. As we predicted, models C-CF-CMF, and F-CF-CMF performed very poorly. This provides support for the conjecture that it is helpful for a developmental system to proceed in an orderly fashion from one scale to a neighboring scale during training.

4 Summary and Conclusions

This paper has considered the hypothesis that systems learning aspects of visual perception may benefit from suitably designed developmental sequences. We reported the results of simulations in which several different systems were trained to detect binocular disparities. One system was non-developmental, in that its data remained the same throughout training. The other models were developmental, their training data changed during the course of training. The developmental models whose stages satisfied the constraints that (1) each training stage consisted of outputs from a single spatial frequency scale, and (2) training proceeded in an orderly fashion from scale to neighboring scale, consistently outperformed both the non-developmental model and the developmental models which violated those constraints. This result supports the notion that the aforementioned constraints are important to learning to detect binocular disparities. We conclude that suitably designed developmental sequences can be useful in the acquisition of binocular disparity sensitivity, in agreement with the “Less is More” hypothesis. Further, that visual development can aid visual learning, and should be further studied.

References

- [1] M. Dominguez and R. A. Jacobs, “Visual development and the acquisition of binocular disparity sensitivities,” in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.
- [2] J. Piaget, *The Origins of Intelligence in Children*, International Universities Press, New York, 1952.
- [3] G. Turkewitz and P. A. Kenney, “Limitations on input as a basis for neural organization and perceptual development: a preliminary statement,” *Developmental Psychobiology*, vol. 15, no. 4, pp. 357–368, 1982.
- [4] E. L. Newport, “Maturational constraints on language learning,” *Cognitive Science*, vol. 14, pp. 11–28, 1990.
- [5] E. L. Newport, “Constraints on learning and their role in language acquisition: Studies of the acquisition of American Sign Language,” *Language Sciences*, vol. 10, pp. 147–172, 1988.
- [6] J. L. Elman, “Learning and development in neural networks: the importance of starting small,” *Cognition*, vol. 43, pp. 71–99, 1993.

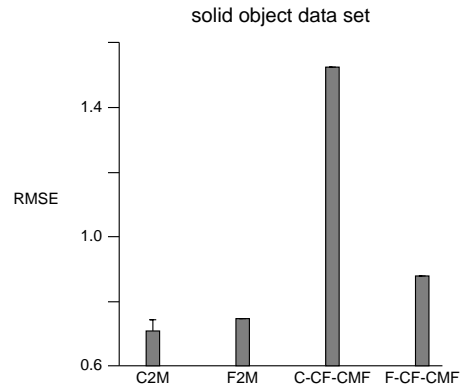


Figure 5: The root mean squared errors (RMSE) of models C2M, F2M, C-CF-CMF, and F-CF-CMF on the test items from the solid object data set.

- [7] D.L.T Rhode and D.C. Plaut, “Language acquisition in the absence of explicit negative evidence: How important is starting small?,” *Cognition*, vol. 72, pp. 67–109, 1999.
- [8] A. Norcia and C. Tyler, “Spatial frequency sweep vep: Visual acuity during the first year of life,” *Vision Research*, vol. 25, pp. 1399–1408, 1985.
- [9] R. Fox, R.N. Aslin, S.L. Shea, and S.T. Dumais, “Stereopsis in human infants,” *Science*, vol. 207, no. 4428, pp. 323–324, 1980.
- [10] R. Held, E. Birch, and J. Gwiazda, “Stereoacuity in human infants,” in *Proceedings of the National Academy of Sciences, USA*, 1980, vol. 77, pp. 5572–5574.
- [11] D. Marr and T. Poggio, “A computational theory of human stereo vision,” in *Proceedings of the Royal Society of London*, 1979, pp. 301–328.
- [12] L.H. Quam, “Hierarchical warp stereo,” Tech. Rep. 402, AI Center, SRI International, 1986.
- [13] S.T. Barnard, “Stereo matching by hierarchical, microcanonical annealing,” Tech. Rep. 414, AI Center, SRI International, 1987.
- [14] H.A. Mallot, S. Gillner, and P.A. Arndt, “Is correspondence search in human stereo vision a coarse-to-fine process?,” *Biological Cybernetics*, vol. 74, pp. 95–106, 1996.
- [15] S.P. McKee and G.J. Mitchison, “The role of retinal correspondence in stereoscopic matching,” *Vision Research*, vol. 28, pp. 1001–1012, 1988.
- [16] P. Mowforth, J.E.W. Mayhew, and J.P. Frisby, “Vergence eye movements made in response to spatial-frequency-filtered random dot stereograms,” *Perception*, vol. 10, pp. 299–304, 1981.
- [17] H.S. Smallman, “Fine-to-coarse scale disambiguation in stereopsis,” *Vision Research*, vol. 34, pp. 2971–2982, 1995.
- [18] M.S. Gray, A. Pouget, R.S. Zemel, S.J. Nowlan, and T.J. Sejnowski, “Reliable disparity estimation through selective integration,” *Visual Neuroscience*, vol. 15, pp. 511–528, 1998.
- [19] I. Ohzawa, G.C. DeAngelis, and R.D. Freeman, “stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors,” *Science*, vol. 249, pp. 1037–1041, 1990.