Kalman Filtering and Neural Networks, Edited by Simon Haykin

ISBNs: 0-471-36998-5 (Hardback); 0-471-22154-6 (Electronic)

3

LEARNING SHAPE AND MOTION FROM IMAGE SEQUENCES

Gaurav S. Patel

Department of Electrical and Computer Engineering, McMaster University, Hamilton, Ontario, Canada

Sue Becker and Ron Racine

Department of Psychology, McMaster University, Hamilton, Ontario, Canada
(beckers@mcmaster.ca)

3.1 INTRODUCTION

In Chapter 2, Puskorius and Feldkamp described a procedure for the supervised training of a recurrent multilayer perceptron – the node-decoupled extended Kalman filter (NDEKF) algorithm. We now use this model to deal with high-dimensional signals: moving visual images. Many complexities arise in visual processing that are not present in one-dimensional prediction problems: the scene may be cluttered with back-

ground objects, the object of interest may be occluded, and the system may have to deal with tracking differently shaped objects at different times. The problem we have dealt with initially is tracking objects that vary in both shape and location. Tracking differently shaped objects is challenging for a system that begins by performing local feature extraction, because the features of two different objects may appear identical locally even though the objects differ in global shape (e.g., squares versus rectangles). However, adequate tracking may still be achievable without a perfect three-dimensional model of the object, using locally extracted features as a starting point, provided there is continuity between image frames.

Our neural network model is able to make use of short-term continuity to track a range of different geometric shapes (circles, squares, and triangles). We evaluate the model's abilities in three experiments. In the first experiment, the model was trained on images of two different moving shapes, where each shape had its own characteristic movement trajectory. In the second experiment, the training set was made more difficult by adding a third object, which also had a unique motion trajectory. In the third and final experiment, the restriction of one direction of motion per shape was lifted. Thus, the model experienced the same shape traveling in different trajectories, as well as different shapes traveling in the same trajectory. Even under these conditions, the model was able to learn to track a given shape for many time steps and anticipate both its shape and location many time steps into the future.

3.2 NEUROBIOLOGICAL AND PERCEPTUAL FOUNDATIONS

The architecture of our model is motivated by two key anatomical features of the mammalian neocortex, the extensive use of feedback connections, and the hierarchical multiscale structure. We discuss briefly the evidence for, and benefits of, each of these in turn.

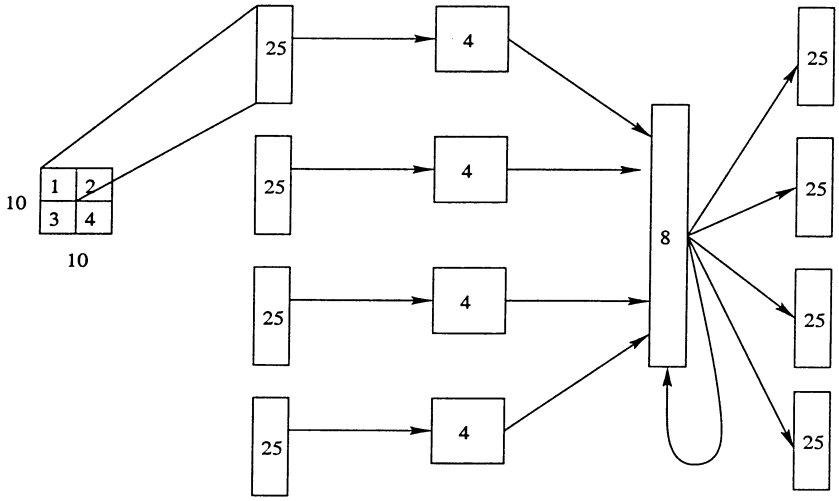
Feedback is a ubiquitous feature of the brain, both between and within cortical areas. Whenever two cortical areas are interconnected, the connections tend to be bidirectional [1]. Additionally, within every neocortical area, neurons within the superficial layers are richly interconnected laterally via a network of horizontal connections [2]. The dense web of feedback connections within the visual system has been shown to be important in suppressing background stimuli and amplifying salient or foreground stimuli [3]. Feedback is also likely to play an important role in processing sequences. Clearly, we view the world as a continuously

varying sequence rather than as a disconnected collection of snapshots. Seeing the world in this way allows recent experience to play a role in the anticipation or prediction of what will come next. The generation of predictions in a perceptual system may serve at least two important functions: (1) To the extent that an incoming sensory signal is consistent with expectations, intelligent filtering may be done to increase the signal-to-noise ratio and resolve ambiguities using context. (2) When the signal violates expectations, an organism can react quickly to such changing or salient conditions by de-emphasizing the expected part of the signal and devoting more processing capacity to the unexpected information. Top-down connections between processing layers, or lateral connections within layers, or both, might be used to accomplish this. Lateral connections allow for local constraints about moving contours to guide one's expectations, and this is the basis for our model.

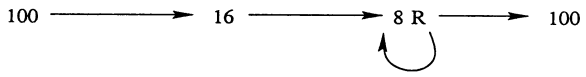
Prediction in a high-dimensional space is computationally complex in a fully connected network architecture. The problem requires a more constrained network architecture that will reduce the number of free parameters. The visual system has done just that. In the earliest stages of processing, cells' receptive fields span only a few degrees of visual angle, while in higher visual areas, cells' receptive fields span almost the entire visual field (for a review, see [4]). Therefore, we designed our model network with a similar hierarchical architecture, in which the first layer of units were connected to relatively small, local regions of the image and a subsequent layer spanned the entire visual field (see Figure 3.1).

3.3 NETWORK DESCRIPTION

Prediction in a high-dimensional space such as a 50×50 pixel image, using a fully connected recurrent network is not feasible, because the number of connections is typically one or more orders of magnitude larger than the dimensionality of the input, and the NDEKF training procedure requires adapting these parameters for typically hundreds to thousands of iterations. The problem requires a more constrained network architecture that will reduce the number of free parameters. Motivated by the hierarchical architecture of real visual systems, we designed our model network with a similar hierarchical architecture in which the first layer of units were connected to relatively small, local 5×5 pixel regions of the image and a subsequent layer spanned the entire visual field (see Figure 3.1).



(*a*)



(*b*)

Figure 3.1 A diagram of the network used. The numbers in the boxes indicate the number of units in each layer or module, except in the input layer, where the receptive fields are numbered $1, \ldots, 4$. Local receptive fields of size 5×5 at the input are fed to the four banks of four units in the first hidden layer. The second layer of eight units then combines these local features learned by the first hidden layer. Note the recurrence in the second hidden layer.

A four-layer network of size 100-16-8R-100, as depicted in Figures 3.1*a* and 3.1*b*, was used in the following experiments. Training images of size 10×10, which are arranged in a vector format of size 100×1, were used to form the input to the networks. As depicted in Figure 3.1*a*, the input image is divided into four non-overlapping receptive fields of size 5×5. Further, the 16 units in the first hidden layer are divided into four banks of four units each. Each of the four units within a bank receive inputs from one of the four receptive fields. This describes how the 10×10 image is connected to the 16 units in the first hidden layer. Each of these 16 units feed into a second hidden layer of 8 units. The second hidden layer has recurrent connections (note that recurrence is only within the layer and not between layers).

Thus, the input layer of the network is connected to small and local regions of the image. The first layer processes these local receptive fields separately, in an effort to extract relevant local features. These features are then combined by the second hidden layer to predict the next image in the sequence. The predicted image is represented at the output layer. The prediction error is then used in the EKF equations to update the weights. This process is repeated over several epochs through the training image sequences until a sufficiently small incremental mean-squared error is obtained.

3.4 EXPERIMENT 1

In the first experiment, the model is trained on images of two different moving shapes, where each shape has its own characteristic movement, that is, shape and direction of movement are perfectly correlated. The sequence of eight 10×10 pixel images in Figure 3.2*a* is used to train a four-layered (100-16-8R-100) network to make one-step predictions of the image sequence. In the first four time steps, a circle moves upward within the image; and in the last four time steps, a triangle moves downward

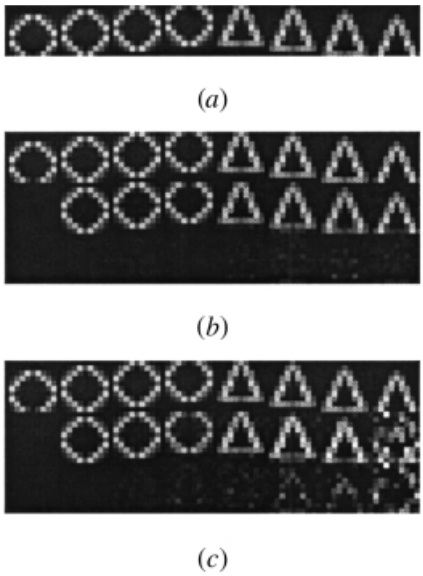


Figure 3.2 Experiment 1: one-step and iterated prediction of image sequence. (*a*) Training sequence used. (*b*) One-step prediction. (*c*) multi-step prediction. In (*b*) and (*c*), the three rows correspond to input, prediction, and error, respectively.

within the image. At each time step, the network is presented with one of the eight 10×10 images as input (divided into four 5×5 receptive fields as described above), and generates in its output layer a prediction of the input at the next time step, but it is always given the correct input at the next time step. Training was stopped after 20 epochs through the training sequence. Figure 3.2*b* shows the network operating in one-step prediction mode on the training sequence after training. It makes excellent predictions of the object shape and also its motion. Figure 3.2*c* shows the network operating in an autonomous mode after being shown only the first image of the sequence. In this multistep prediction case, the network is only given external input at the first time step in the sequence. Beyond the first time step, the network is given its prediction from time $t - 1$ as its input at time t, which could potentially lead to a buildup of prediction errors over many time steps. This shows that the network has reconstructed the entire dynamics, to which it was exposed during training, when provided with only the first image. This is indeed a difficult task. It is seen that as the iterative prediction proceeds, the residual errors (the third row in Figure 3.2*c*) are amplified at each step.

3.5 EXPERIMENT 2

Next, an ND-EKF network with the same 100-16-8R-100 architecture used in Experiment 1 was trained with three sequences, each consisting of four images, in the following order:

- circle moving right and up;
- triangle moving right and down;
- square moving right and up.

During training, at the beginning of each sequence, the network states were initialized to zero, so that the network would not learn the order of presentation of the sequences. The network was therefore expected to learn the motions associated with each of the three shapes, and not the order of presentation of the shapes.

During testing, the order of presentation of the three sequences varied, as shown in Figure 3.3*a*. The trained network does well at the task of one-step prediction, only failing momentarily at transition points where we switch between sequences. It is important to note that one-step prediction, in this case, is a difficult and challenging task because the network has to determine (1) what shape is present and (2) which direction it is moving in, without direct knowledge of inputs some time in the past. In order to

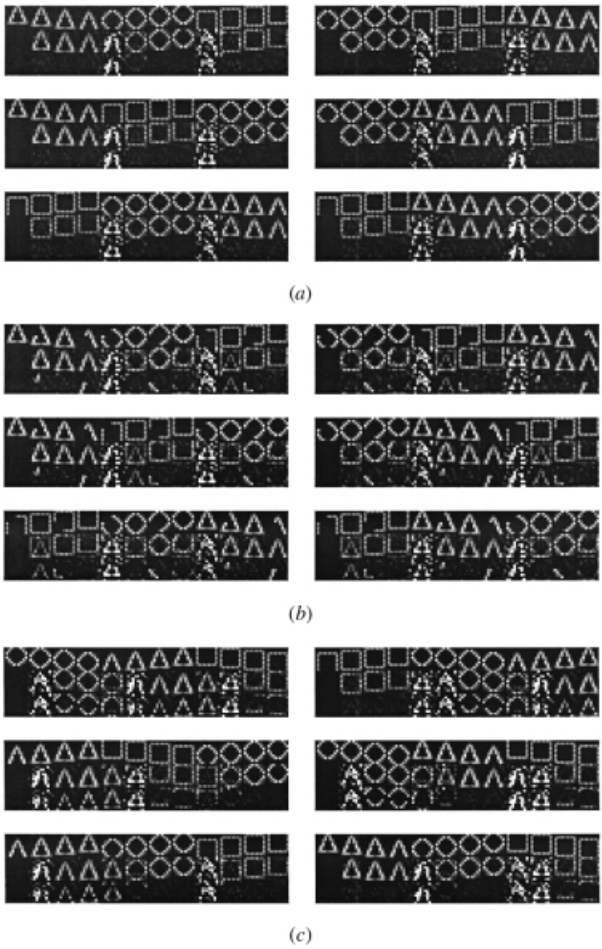


Figure 3.3 Experiment 2: one-step prediction of image sequences using the trained network. (*a*) Various combinations of sequences used in training. (*b*) Same sequences as in (*a*), but with occlusions. (*c*) Prediction on some sequences not seen during training. The three rows in each image correspond to input, prediction, and error, respectively.

make good predictions, it must rely on its recurrent or feedback connections, which play a crucial role in the present model.

We also tested the model on a set of occluded images – images with regions that are intentionally zeroed. Remarkably, the network makes correct one-step predictions, even in the presence of occlusions as shown in Figure 3.3*b*. In addition, the predictions do not contain occlusions; that is, they are correctly filled in, demonstrating the robustness of the model to occlusions. In Figure 3.3*c*, when the network is presented with

sequences that it had not been exposed to during training, a larger residual error is obtained, as expected. However, the network is still capable of identifying the shape and motion, although not as accurately as before.

3.6 EXPERIMENT 3

In Experiment 1, the network was presented with short sequences (four images) of only two shapes (circle and triangle), and in experiment 2 an extra shape (square) was added. In Experiment 3, to make the learning task even more challenging, the length of the sequences was increased to 10 and the restriction of one direction of motion per shape was lifted. Specifically, each shape was permitted to move right and either up or down. Thus, the network was exposed to different shapes traveling in similar directions and also the same shape traveling in different directions, increasing the total number of images presented to the network from 8 images in Experiment 1 and 12 images in Experiment 2 to 100 images in this experiment. In effect, there is a substantial increase in the number of learning patterns, and thus a substantial increase in the complexity of the learning task. However, since the number of weights in the network is limited and remains the same as in the other experiments, the network cannot simply memorize the sequences.

We trained a network of the same 100-16-8R-100 architecture on six sequences, each consisting of 10 images (see Fig. 3.4) in the following order:

- circle moving right and up;
- square moving right and down;
- triangle moving right and up;
- circle moving right and down;
- square moving right and up;
- triangle moving right and down.

Training was performed in a similar manner as Experiment 2. During testing, the order of presentation of the six sequences was varied; several examples are shown in Figure 3.5. As in the previous experiments, even with the larger number of training patterns, the network is able to predict the correct motion of the shapes, only failing during transitions between shapes. It is able to distinguish between the same shapes moving in different directions as well as different shapes moving in the same direction, using context available via the recurrent connections.

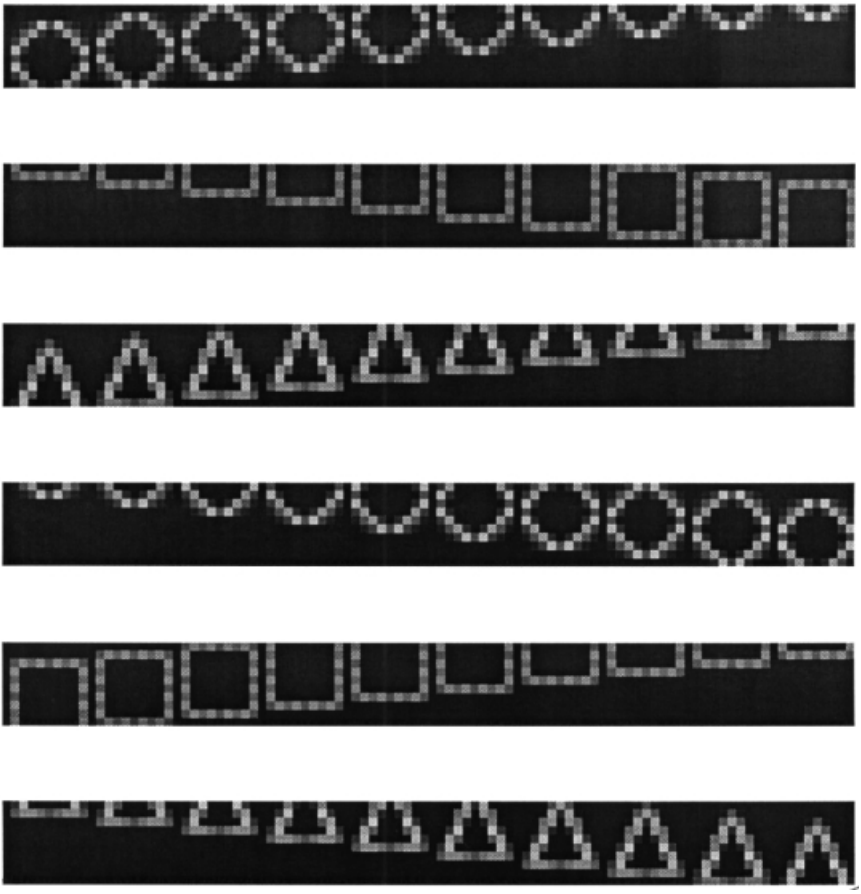


Figure 3.4 Experiment 3: six image sequences used for training.

The failure of the model to make accurate predictions at transitions between shapes can also be seen in the residual error that is obtained during prediction. The residual error in the predicted image is quantified by calculating the mean-squared prediction error, as shown in Figure 3.6. The figure shows how the mean-squared prediction error varies as the prediction continues. Note the transient increase in error at transitions between shapes.

3.7 DISCUSSION

In this chapter, we have dealt with time-series prediction of high-dimensional signals: moving visual images. This situation is much more

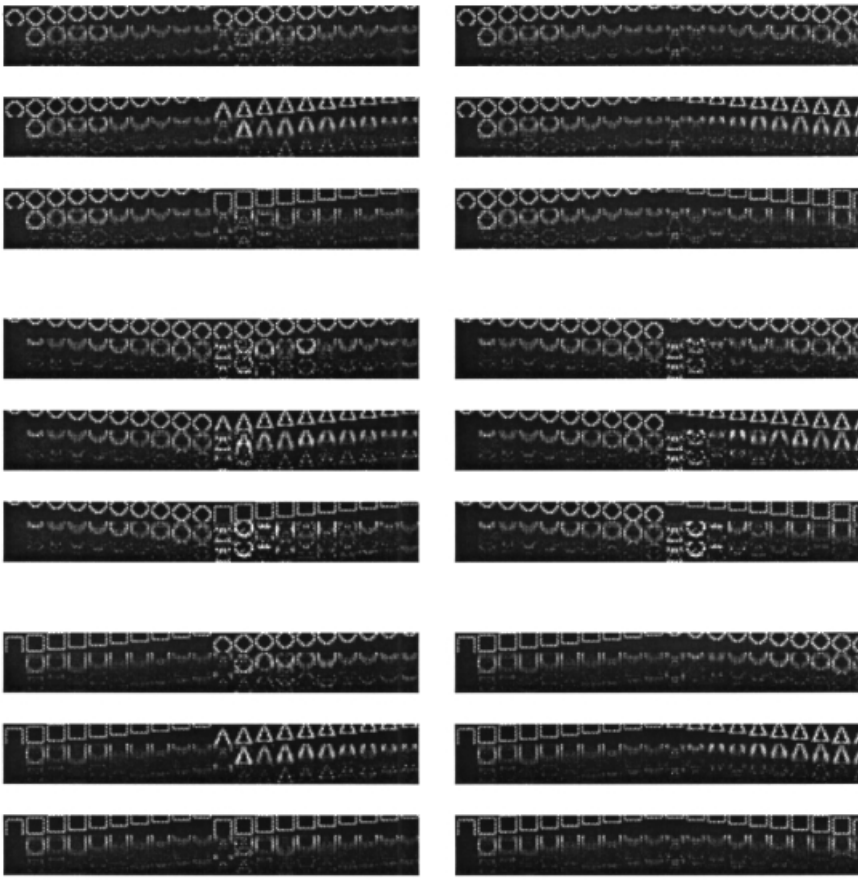


Figure 3.5 Experiment 3: one-step prediction of image sequences using the trained network. The three rows in each image correspond to input, prediction, and error, respectively.

complicated than a one-dimensional case, in that the system has to deal with simultaneous shape and motion prediction. The network was trained by the EKF method to perform one-step prediction of image sequences in a specific order. Then, during testing, the order of the sequences was varied and the network was asked to predict the correct shape and location of the next image in the sequence. The complexity of the problem was increased from Experiment 1 to 3 as we introduced occlusions, increased both the length of the training sequences and the number of shapes presented, and allowed shape and motion to vary independently. In all

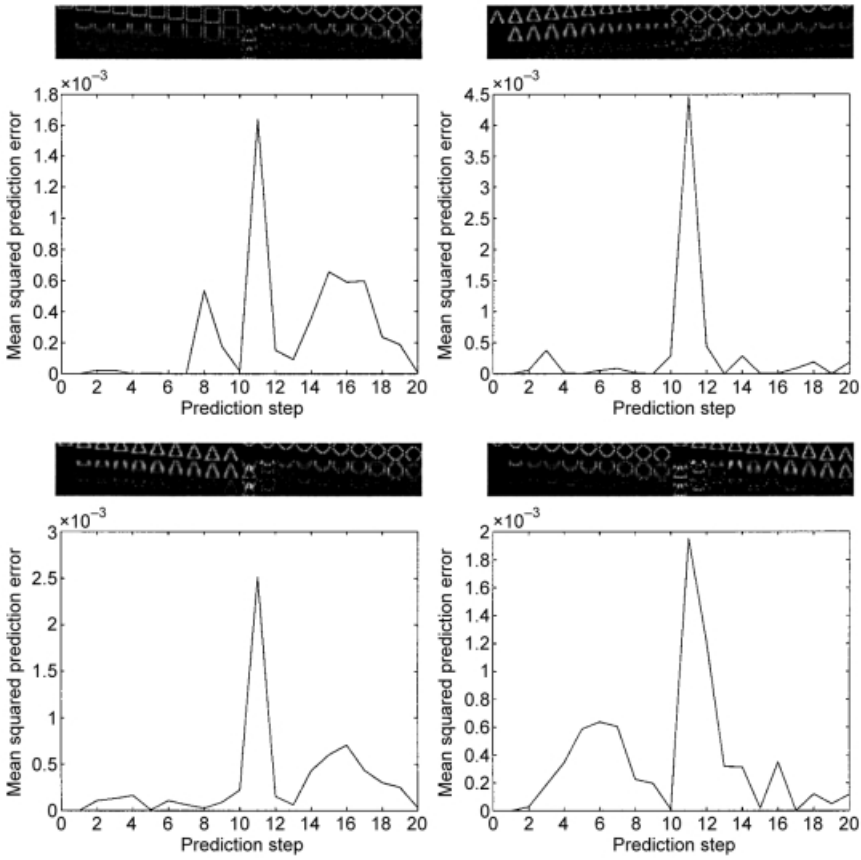


Figure 3.6 Mean-squared prediction error in one-step prediction of image sequences using the trained network. The three rows in each image correspond to input, prediction, and error, respectively. The graphs show how the mean-squared prediction error varies as the prediction progresses. Notice the increase in error at transitions between shapes.

cases, the network was able to predict the correct motion of the shapes, failing only momentarily at transitions between shapes.

The network described here is a first step toward modeling the mechanisms by which the human brain might simultaneously recognize and track moving stimuli. Any attempt to model both shape and motion processing simultaneously within a single network may seem to be at odds with the well-established finding that shape and spatial information are processed in separate pathways of the visual system [5]. An extreme version of this view posits that form-related features are processed strictly by the ventral "what" pathway and motion features are processed strictly

by the dorsal "where" pathway. Anatomically, however, there are cross-connections between the two pathways at several points [6]. Furthermore, there is ample behavioral evidence that the processes of shape and motion perception are not completely separate. For example, it has long been established that we are able to infer shape from motion (see e.g., [7]). Conversely, under certain conditions, object recognition can be shown to drive motion perception [8]. In addition, Stone [9] has shown that viewers are much better at recognizing objects when they are moving in characteristic, familiar trajectories as compared with unfamiliar trajectories. These data suggest that when shape and motion are tightly correlated, viewers will learn to use them together to recognize objects. This is exactly what happens in our model.

To accomplish temporal processing in our model, we have incorporated within-layer recurrent connections in the architecture used here. Another possibility would be to incorporate top-down recurrent connections. A key anatomical feature of the visual system is top-down feedback between visual areas [3]. Top-down connections could allow global expectations about the three-dimensional shape of a moving object to guide predictions. Thus, an important direction for future work is to extend the model to allow top-down feedback. Rao and Ballard [10] have proposed an alternative neural network implementation of the EKF that employs top-down feedback between layers, and have applied their model to both static images and time-varying image sequences [10, 11]. Other models of cortical feedback for modeling the generation of expectations have also been proposed (see, e.g., [12, 13]).

Natural visual systems can deal with an enormous space of possible images, under widely varying viewing conditions. Another important direction for future work is to extend our model to deal with more realistic images. Many additional complexities arise in natural images that were not present in the artificial image sequences used here. For example, the simultaneous presence of both foreground and background objects may hinder the prediction accuracy. Natural visual systems likely use attentional filtering and binding strategies to alleviate this problem; for example, Moran and Desimone [14] have observed cells that show a suppressed neural response to a preferred stimulus if unattended and in the presence of an attended stimulus. Another simplification of our images is that shape remained constant for many time frames, whereas for real three-dimensional objects, the shape projected onto a two-dimensional image may change dramatically over time, because of rotations as well as non-rigid motions (e.g. bending). Humans are able to infer three-dimensional shape from non-rigid motion, even from highly impoverished stimuli such

as moving light displays [7]. It is likely that the architecture described here could handle changes in shape, provided shape changes predictably and gradually over time.

REFERENCES

[1] D.J. Felleman and D.C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex", *Cerebral Cortex*, **1**, 1–47 (1991).

[2] J.S. Lund, Q. Wu and J.B. Levitt, "Visual cortex cell types and connections", in M.A. Arbib, Ed., *Handbook of Brain Theory and Neural Networks*, Cambridge, MA: MIT Press, 1995.

[3] J.M. Hupé, A.C. James, B.R. Payne, S.G. Lomber, P. Girard and J. Bullier, "Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons", *Nature*, **394**, 784–787 (1998).

[4] M.W. Oram and D.I. Perrett, "Modeling visual recognition from neurobiological constraints", *Neural Networks*, **7**, 945–972 (1994).

[5] M. Mishkin, L.G. Ungerleider and K.A. Macko, "Object vision and spatial vision: Two cortical pathways", *Trends in Neurosciences*, **6**, 414–417 (1983).

[6] E.A. De Yoe and D.C. Van Essen, "Concurrent processing streams in monkey visual cortex", *Trends in Neurosciences*, **11**, 219–226, (1988).

[7] G. Johanssen, "Visual perception of biological motion and a model for its analysis", *Perception and Psychophysics*, **14**, 201–211 (1973).

[8] V.S. Ramachandran, C. Armel, C. Foster and R. Stoddard, "Object recognition can drive motion perception", *Nature*, **395**, 852–853 (1998).

[9] J.V. Stone, "Object recognition: View-specificity and motion-specificity", *Vision Research*, **39**, 4032–4044, (1999).

[10] R.P.N. Rao and D.H. Ballard, "Dynamic model of visual recognition predicts neural response properties in the visual cortex", *Neural Computation*, **9(4)**, 721–763 (1997)

[11] R.P.N. Rao, "Correlates of attention in a model of dynamic visual recognition", in M.I. Jordan, M.J. Kearns and S.A. Solla, Eds., *Advances in Neural Information Processing Systems*, Vol. 10. Cambridge, MA: MIT Press, 1998.

[12] E. Harth, K.P. Unnikrishnan and A.S. Panday, "The inversion of sensory processing by feedback pathways: A model of visual cognitive functions", *Science*, **237**, 184–187 (1987).

[13] D. Mumford, "On the computational architecture of the neocortex", *Biological Cybernetics*, **65**, 135–145 (1991).

[14] J. Moran and R. Desimone, "Selective attention gates visual processing in the extrastriate cortex", *Science*, **229**, 782–784, (1985).