Max-Planck-Institut
für biologische Kybernetik
Arbeitsgruppe Bülthoff

Spemannstraße 38 • 72076 Tübingen • Germany

Technical Report No. 37 August 1996

How neurons learn to associate 2D-views in invariant object recognition

Guy Wallis

Abstract

A local learning rule is shown to be able to account for the association of images together on the basis of temporal order rather than spatial configuration, as described in single cell recording results published by Miyashita (1988). Possible reasons for requiring such learning are then given in the context of invariant object recognition

This work was supported by a Fellowship from the Max-Plank Gesellschaftschaft

1 Introduction

The primate temporal lobe has long been implicated in the recognition of objects (Ungerleider & Mishkin, 1982; Goodale & Milner, 1992). Single cell recording in this area has revealed neurons responsive to images of faces (Desimone, 1991; Rolls, 1992) and other, more abstract image features (Tanaka et al., 1991; Miyashita & Chang, 1988). Of particular interest is the ability of these cells to demonstrate a robustly invariant response to a preferred stimulus object as it undergoes large rigid transformations in size, rotation in depth and location. This raises the question of how a single neuron might learn to associate spatially very dissimilar images together as belonging to the same object. Any approach based solely upon physical appearance cannot hope to capture all of the invariances which have been described in the literature, one of which forms the basis for the learning described and implemented here.

The experiment in question was carried out by Miyashita (1988) in which randomly generated colour fractal patterns were presented to macaque monkeys. The animals' task was to observe a pattern and then indicate whether a subsequent pattern was the same or not. Testing proceeded from trial to trial with a consistent order of testing being maintained throughout. Examples of the types of patterns used appear in figure 1 and an overview of the testing regime appears in figure 3. Although the experimental paradigm did not explicitly require the overall test sequence to be remembered, Miyashita discovered that neurons within inferior temporal lobe became responsive to an ordered subset of the 96 images in the test set.

The fact that the images were generated randomly meant that there was no particular reason - on the grounds of spatial similarity - why these images should have become associated together by a single neuron. Instead the results indicate the importance of the temporal order in controlling the learning of neural selectivity.

In this work I shall be presenting a paradigm in which neurons are able to associate images appearing in spatio-temporal sequences from a set of images in the manner which Miyashita has described, explaining why his results may be a special case of a more general learning mechanism. I shall also compare the work to earlier research on the topic by Griniasty, Tsodyks & Amit (1993) as well as other recent work on object recognition. Since the neurons which Miyashita studied lie in temporal lobe, temporal order based association learning I will finish by describing possible important consequences for invariant object recognition.

2 Learning from spatio-temporal associations

2.1 Introduction

The first aim of this paper is to propose a biologically justifiable learning mechanism and neural architecture for implementing the temporal aspect of the learning which Miyashita describes.

One interpretation of Miyashita's results is that they can be implemented in the dynamics of an attractor neural network in which attractor states of patterns close in temporal presentation order overlap (Griniasty et al., 1993). Although this work did not include simulation results it did describe how the network of connections required might be set up with some form of time averaging filter acting either on the inputs or outputs of each neuron. The applicability of Hebbian dynamics which Griniasty *et al.* described is echoed in the simulations described here, although I will propose a much simpler neural architecture capable of producing the time based correlations sought and will describe reasons for questioning their recurrent processing model.

2.2 Learning rule

A suitable time averaging learning rule has already been described with reference to object recognition (Földiák, 1991; Wallis & Rolls, 1996). The learning rule works by replacing the current neural activation term found in standard Hebbian learning paradigms, with a running, time averaged measure - called the 'trace' value.

The trace learning rule used is equivalent to Földiák's (1991) , and can be summarised as follows:

$$\Delta {w_{ij}}^{(t)} = \alpha {\overline{y}_i}^{(t)} . x_j \quad : \quad \sum_j {w_{ij}}^2 = 1$$

and

$${\overline{y}_i}^{(t)} = (1-\eta) {y_i}^{(t)} + \eta {\overline{y}_i}^{(t-1)} \; : \; y_i = \Phi \left[\sum_j x_j w_{ij} \right]$$

where x_j is the j^{th} input to the neuron, y_i is the output of the i^{th} neuron, w_{ij} is the j^{th} weight on the i^{th} neuron, $\eta = 0.6$ governs the relative influence of the trace and the new input, and ${\overline{y}_i}^{(t)}$ represents the value of the i^{th} cell's trace at time t. The function Φ implements a non-linear activation function as well as local inhibition details of which are described in the next section.

The essence of how object invariance might be learned with such a rule can be seen by considering the situation in which a single neuron is strongly activated by some element of a real world object. In such a case the short-term average activity of

the neuron will be high, and if a new aspect of the object is seen before the effects of this activity die away - in the order of 0.5s - then not only will the initially active afferent synapses modify onto the neuron, but so also will the synapses activated by the transformed version of this stimulus. In this way the cell will learn to respond to either appearance of the original stimulus. Making such associations works in practice because it is very likely that within short time periods different aspects of the same object will be being inspected. The cell will not, however, tend to make spurious links across stimuli that are part of different objects because of the unlikelihood in the real world of one object consistently following another.

Several means for implementing this learning rule in real neurons have been described (Wallis & Rolls, 1996), however one of particular relevance here makes use of the prolonged firing of temporal lobe neurons for 100–200ms even after very rapid presentations of an effective stimulus (Rolls & Tovee, 1994). It is suggested that this would, in natural circumstances, be time enough for new views of the effective stimulus object to be seen and learnt. This supposes that firing at the soma should not only propagate along the axon but also be capable of affecting learning in the dendritic tree. Evidence to support this claim has in fact recently been reported in rat neocortical layer V pyramidal neurons (Markram et al., 1995).

Before proceeding it is important to discuss a problem with applying this form of learning paradigm to the case studied by Miyashita. Making associations between stimuli over the long delay period of 16s which he describes, would not normally be desirable - since the viewer would typically have moved his attention to a new object. This in turn might lead to the spurious linking of objects mentioned above. In order to explain this, it is worth considering Miyashita & Chang's earlier paper (1988) in which they explicitly describe greatly extended periods of maintained firing throughout the delay period which would in fact allow learning to proceed by the associative mechanism described above.

Under normal viewing conditions neural activity only proceeds for a few hundred milliseconds after the removal of the activating stimulus (Rolls & Tovee, 1994). So why might neurons be firing for so long in experiments reported by Miyashita and Chang? There is now reasonable evidence that the delayed match to sample (DMS) paradigm used in Miyashita's experiments represent a rather special case in image analysis. It seems that the animals' solution to the DMS experiment involves maintaining activity in selective cells during the delay period. This maintenance of activity is itself probably mediated by neurons outside the temporal lobe, in prefrontal cortex (Desimone et al., 1995; Fuster et al., 1985). In addition, even in the case of a DMS experiment, the appearance of other images during the delay period quickly abolishes any maintained activity (Baylis & Rolls, 1987; Miller & Desimone, 1994). In other words, under normal viewing conditions associations would not be made over the large delay periods used in Miyashita's experiments. However, if the memory of the activity of the neuron is explicitly maintained then such associations can indeed be made.

The absence of long periods of maintained activity under normal viewing conditions calls into question the general applicability of the recurrent processing model proposed by Griniasty *et al.* Further reasons for preferring the feed-forward architecture used here - on the grounds of speed of processing and absence of gradual adaptation in neural activity - have also been described (Rolls & Tovee, 1994; Thorpe & Imbert, 1989).

2.3 Network architecture

A two layer network was constructed - see figure 2. The first layer acts as a local feature extraction layer and consists of a three (one per colour channel) 32x32 grids of neurons arranged in 64 4x4 inhibitory pools. Each pool fully samples a corresponding 4x4 patch of the 32x32 input image. Competition within these pools is of the 'winner take *most*' type, otherwise referred to as leaky learning (Hertz et al., 1990). In the context of this network, this implies establishing which neuron within each pool is firing most strongly and electing it the winner. All other neurons within the same pool then have their firing rate reduced to one third of their initial rate so as to implement local inhibition. All learning in this layer is simple Hebbian.

Above the input layer is a second layer consisting of a single inhibitory pool of 16 neurons which fully samples the first layer. Neurons in this layer are trained with the trace rule. All neurons in both layers also have a separate, non-linear activation function which transforms the cell's calculated weighted input into an output firing rate. This was achieved by scaling the outputs within each inhibitory pool to 1 and then passing the result through a sigmoidal activation function. The action of the inhibition and non-linear activation function are represented by the function Φ in the previous trace rule equations. The rescaling was intended to keep the amount of learning taking place for each stimulus roughly constant.

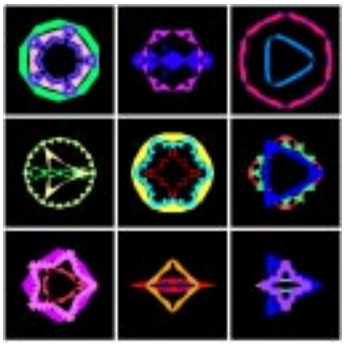

Figure 1: Nine example fractal images.

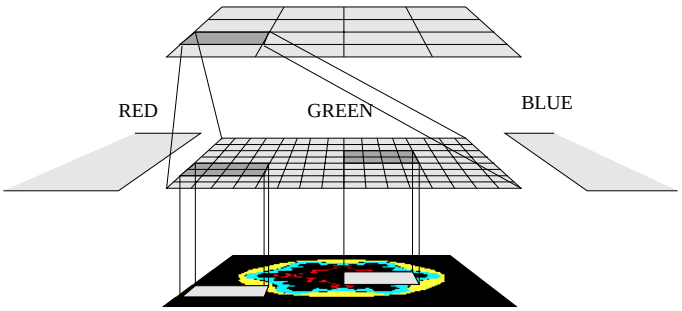

Figure 2: The two layer network used in the simulations, containing three separate input layers, one for each colour channel.

3 Simulating fractal image learning

3.1 Introduction

This section describes a series of experiments carried out to discover whether the results described Miyashita could be implemented using the network and learning rule described above.

3.2 Methods

The stimuli presented during training appear in figure 1. They were generated using the algorithm used described by Miyashita *et al.* (1991). A stimulus was presented to the network long enough for the running average activity (the trace value, $\overline{y}$ in the earlier equations) to saturate. This activity was then maintained at this level during the delay period. A second stimulus was then presented which was either the same as the first stimulus, or chosen at random from the 96 other images, both with probability 0.5. Activity was similarly maintained onto the next trial in which the next image in the set sequence was trained. After training the network for 800 complete cycles the net was tested on both the original training set and a further 97 novel fractal images generated by the same algorithm.

3.3 Results

Figure 4 show the responses of two cells to the 97 trained stimuli as well as the 97 novel stimuli. The bunching of strong responses along the image number axis clearly demonstrate the preference of these neurons for groups of stimuli which appeared closely in time. The more sporadic form of the responses to the novel images also confirms that this is an effect of learning. These results are in general agreement with the form of curves plotted in Miyashita's paper (1988). A more direct comparison is provided by the autocorrelation function for these responses, which appear averaged over all 16 output cells in figure 5. The smoothly decaying curve seen for the learned stimuli demonstrates a strong correlation between responses to neighbouring images in the sequence and is in stark contrast to the correlation for responses to the novel stimuli. Correlation becomes indistinguishable from zero at around five image steps away from the the central stimulus, a figure also in close accord with those provided in Miyashita's paper.

4 Conclusions

This paper has demonstrated that a local Hebb-like learning rule can train neurons to associate images appearing in time, in accordance with single cell recording data described by Miyashita (1988) . By reproducing Miyashita's results, the earlier work of both Foldiak and Wallis concerning the use of a time based learning rule in object recognition (Földiák, 1991; Wallis & Rolls, 1996; Wallis, 1996) , has found strong support in data from real neurophysiological recordings.

I have argued that a reason for associating images on the basis of consistent correlations in their appearance in time would be to help solve the problem of invariant object recognition. As a corollary to this, it might be possible for humans to generalise between objects if views of them appear in artificial repetitive sequences. This type of over-generalisation forms the basis for psychophysical work currently underway.

References

Baylis, G. and Rolls, E. (1987). Responses of neurons in the inferior temporal cortex in short term and serial recognition memory tasks. *Experimental Brain Research*, 65:614–622.

Desimone, R. (1991). Face–selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience*, 3:1–8.

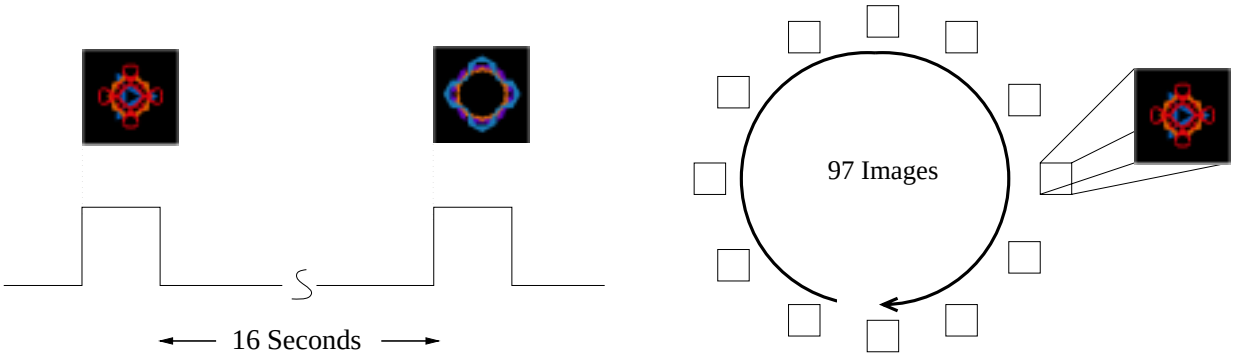

Figure 3: Overview of the testing paradigm used by Miyashita (1988) showing the presentation timing and repeating sequence of 97 fractal image test stimuli.

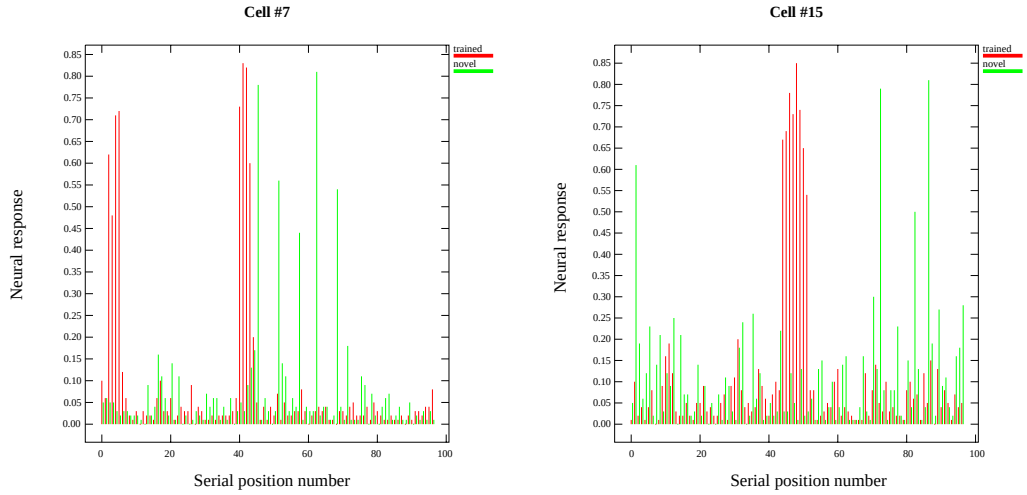

Figure 4: The response of two neurons to all 97 test images and 97 novel fractal images. The contrast in the degree of clustering and amplitude of responses demonstrates effective learning.

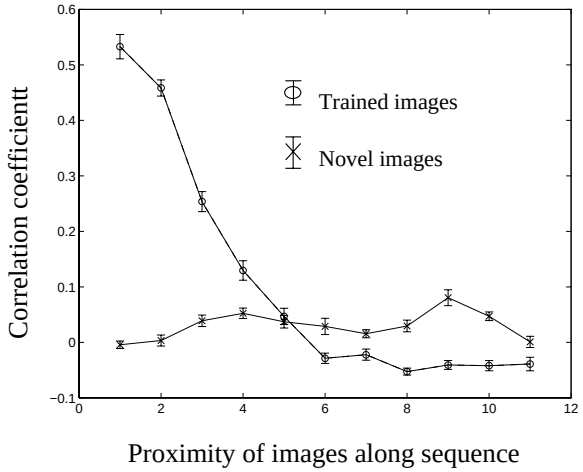

Figure 5: Average autocorrelation function for the responses of all 16 cells to the trained and novel test sets.

Desimone, R., Miller, E., Chelazzi, L., and Lueschow, A. (1995). Multiple memory systems in visual cortex. In Gazzaniga, M., editor, *Cognitive Neurosciences*, chapter 30, pages 475–486. MIT Press: New York.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3:194–200.

Fuster, J., Bauer, R., and Jervey, J. (1985). Functional interactions between inferotemporal and prefrontal cortex in a cognitive task. *Brain Research*, 330:299–307.

Goodale, M. and Milner, A. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15:20–25.

Griniasty, M., Tsodyks, M., and Amit, D. (1993). Conversion of temporal correlations between stimuli to spatial correlations between attractors. *Neural Computation*, 35:1–17.

Hertz, J., Krogh, A., and Palmer, R. (1990). *Introduction to the theory of neural computation.* Santa Fe Institute: Addison Wesley.

Markram, H., Helm, P., and Sakmann, B. (1995). Dendritic calcium transients evoked by single back-propagating action potentials in rat neocortical pyramidal neurons. *Journal of Physiology*, 485:1–20.

Miller, E. and Desimone, R. (1994). Parallel neuronal mechanisms for short-term memory. *Science*, 254:1377–1379.

Miyashita, Y. (1988). Neuronal correlate of visual associative long–term memory in the primate temporal cortex. *Nature*, 335:817–820.

Miyashita, Y. and Chang, H. (1988). Neuronal correlate of pictorial short–term memory in the primate temporal cortex. *Nature*, 331:68–70.

Miyashita, Y., Higuchi, S., Sakai, K., and Masui, N. (1991). Generation of fractal patterns for probing the visual memory. *Neuroscience Research*, 12:307–311.

Rolls, E. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical areas. *Philosophical Transactions of the Royal Society, London [B]*, 335:11–21.

Rolls, E. and Tovee, M. (1994). Processing speed in the cerebral cortex, and the neurophysiology of visual masking. *Proceedings of the Royal Society, London [B]*, 257:9–15.

Tanaka, K., Saito, H., Fukada, Y., and Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, 66:170–189.

Thorpe, S. and Imbert, M. (1989). Biological constraints on connectionist models. In Pfeifer, R., Schreter, Z., and Fogelman-Soulie, F., editors, *Connectionism in Perspective*, pages 63–92. London: John Wiley and Sons.

Ungerleider, L. and Mishkin, M. (1982). Two cortical visual systems. In Ingle, D., Goodale, M., and Mansfield, R., editors, *Analysis of Visual Behaviour*, pages 549–586. Cambridge, Massachusetts, USA: MIT press.

Wallis, G. (1996). Using spatio-temporal corrolations to learn invariant object recognition. *To appear*
in Neural Networks. www: ftp://ftp.mpik-tueb.mpg.de/pub/guy/nn.ps.Z.

Wallis, G. and Rolls, E. (1996). A model of invariant object recognition in the visual system. *Progress in Neurobiology, submitted for review.* www: ftp://ftp.mpik-tueb.mpg.de/pub/guy/pnb.ps.Z.