Using Spatio-Temporal Correlations to Learn Invariant Object Recognition

Guy Wallis Max-Planck Institute für biologische Kybernetik, Spemannstraße 38, 72076 Tübingen, Germany. Email: guy@mpik-tueb.mpg.de

A competitive network is described which learns to classify objects on the basis of temporal as well as spatial correlations. This is achieved by using a Hebb-like learning rule which is dependent upon prior as well as current neural activity. The rule is shown to be capable of outperforming a supervised rule on the cross-validation test of an invariant character recognition task, given a relatively small training set. It is also shown to outperform the supervised version of Fukushima's Neocognitron (1980), on a larger training set.

1 Introduction

Competitive networks have been successfully applied to many classification problems in which specifications of the stimuli are encoded as a stimulus description vector and in which the network groups these descriptors on the basis of their Euclidean separation (Hertz *et al.*, 1990; Rumelhart *et al.*, 1986; Hecht-Nielsen, 1990, for review). When faced with the problem of categorizing images of stimuli which have been deformed by simple transformations such as translation, these descriptors must be carefully chosen if they are not to appear very different as inputs to the network. Numerous ideas including weight sharing have been proposed to accommodate for such translations (Fukushima, 1980, for example). This paper describes simulations of an unsupervised learning rule which requires information about the activity within a local region, but which otherwise acts locally within a cell, hence making it a more biologically relevant alternative. The learning rule itself uses spatio-temporal correlations of stimuli as well as spatial ones to assemble stimulus categories, building on the assumption that images of objects appear in short sequences - such as we experience in everyday life by attending to and observing objects. This learning rule, henceforth referred to as the trace rule, has received some attention previously (Földiák, 1991; Wallis *et al.*, 1993; Wallis & Rolls, 1995), but the aim of this paper is to provide the first comparative study of the rule in relation to other learning rules and architectures.

The version of the trace rule used in this work is equivalent to Földiák's[1], and

[1] Other researchers have also described the use of such a learning scheme in classical conditioning (Klopf, 1988; Sutton & Barto, 1981).

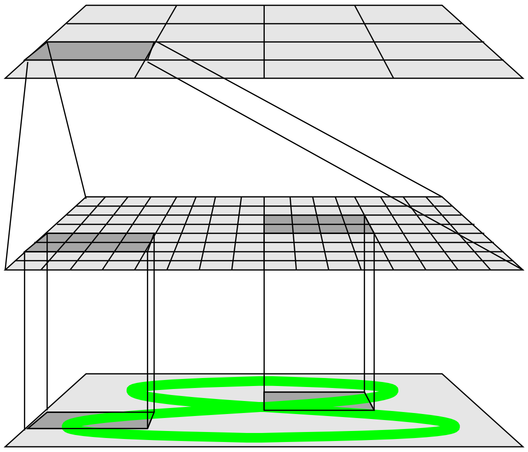

Figure 1: The TraceNet network architecture.

can be summarized as follows:

$$\Delta w_{ij}{}^{(t)} = \alpha \overline{y}_i{}^{(t)}.x_j \quad : \quad \sum_j w_{ij}{}^2 = 1 \; for \; each \; i^{th} \; neuron$$

and

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

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, η 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 factor k implements a non-linear activation function for each neuron, and the function Φ implements local lateral inhibition within a local region of neurons as well as a renormalisation of local activity, as explained in the following experiments. Note that no explicit resetting of the trace value between presentation sequences is required or used.

This paper sets out to test the performance of the trace rule by implementing it in a simple two-layer competitive network and comparing the overall performance achieved with that of a supervised learning rule, and other established architectures. To do this, two hand-written character code recognition tasks are performed.

2 Learning Stimulus Classes Using the Trace Rule and Delta Rule

2.1 Experiment I: Classification Performance of the Trace Rule and the Delta Rule

This first experiment uses a classical handwritten digit classification task to establish the ability of the unsupervised trace rule to categorize the digits presented, and will also serve to verify the use of a simple two-layer network used in the simulations. In addition, in order to gauge the relative performance of the trace rule, the same network is also trained using the standard delta rule (Widrow & Hoff, 1960).

The two layer network, referred to as the TraceNet was constructed with the intention that the first layer should serve to extract some local features which might appear comparable in transformed versions of the same digit. The first layer consists of 256 neurons arranged in a 4x4 grid of 4x4 inhibitory pools. Each pool fully samples a corresponding 4x4 patch of a 16x16 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, before normalising all activity within a pool to a fixed maximum of one. This process implements the local inhibition and is represented as 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. All learning in this layer is simple Hebbian.

Centered above this input layer is a second, output layer consisting of a single, similarly laterally connected pool of ten neurons all fully connected to the 256 neurons in the first layer. The neurons in this layer are trained with either the trace rule or delta rule - as described in the following experiments. Figure 1 shows the general architecture.

All of the network neurons 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 raising the activation value to a fixed power, set at 7 and represented as k in the trace rule equations. This non-linearity served to reduce the overall activity within a layer and hence direct learning onto the few remaining active neurons. The value 7 was found to be the optimal integer value in the range 1 to 10, optimally sparsifying activity within the network whilst stopping short of deactivating all other neurons within an inhibitory pool (Wallis, 1994).

As a test of recognition performance, digits from a character set used by LeCun *et. al.* (1989) were presented to the network. During learning, 100 digits - 10 of each type - were presented in permuted random sequence, with the only constraint being that all ten examples of a particular digit were shown in sequence, once that digit was chosen to be presented, thereby affording the temporal structure

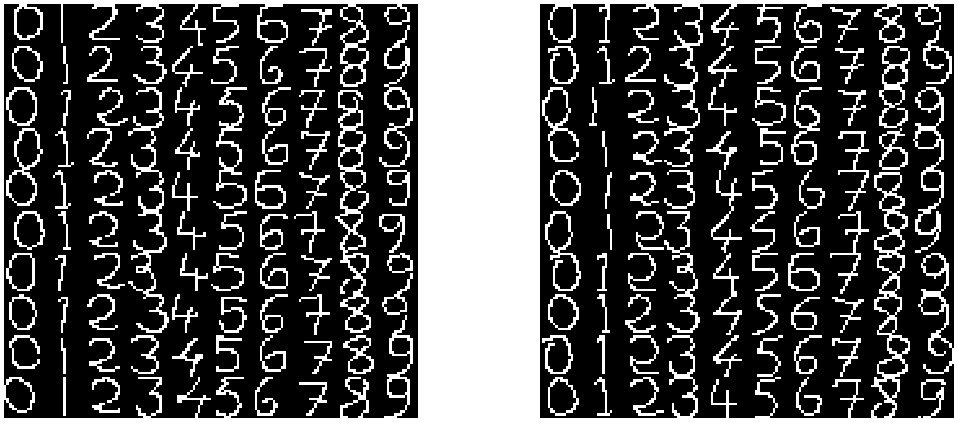

Figure 2: The training set (left) and cross validation data set (right), taken from the LeCun (1989) data set.

necessary for operation of the trace rule. Cross validation testing was conducted by presenting a further 100 digits from the same database - see figure 2. The task for the learning rules was, then, to demonstrate their ability to solve the invariance problem posed - i.e. to categorize digits from each class together, and further, to yield good categorization performance on the cross-validation data set.

2.2 Experiment I: Results

The results of training with the two training signals are given in figure 3 - each data point representing the average performance taken over five runs of the network, with performance being rated as the percentage fraction of transmitted information about stimulus class identity seen in the network output (see e.g.(Attneave, 1959)). The value of the trace rule parameter η was set at 0.8 in this experiment[2]. The trace rule is clearly highly effective, in that it achieves rapid convergence to near optimal performance - achieving results as high as 95%, almost equaling the final performance of the delta rule, which correctly categorises the whole training set.

More importantly however, the cross-validation performance achieved by the trace rule is much improved over the rather poor results achieved with the delta rule. Due to the relatively small training set, the delta rule is seen to overfit the data, finding a somewhat arbitrary linear combination of inputs to solve the specific training problem. The trace rule has, however, found a more robust solution, the reasons for which are explored in the following section.

As an aside, performance was also tested with η set to zero. This corresponds to simple Hebbian learning in both the first and now also the second layer - for which

[2]The relationship between the value of η and the length of stimulus presentation has been explored elsewhere (Wallis, 1994).

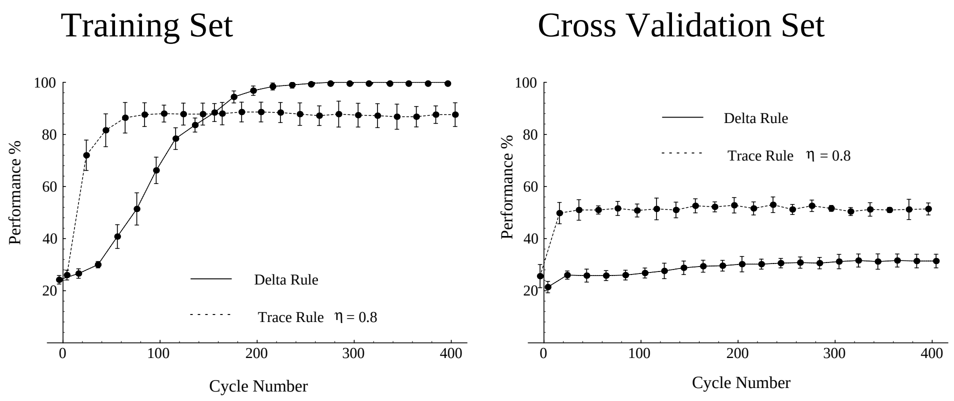

Figure 3: Trace Rule vs Delta Rule. The left hand graph shows the network performance on the training set. The right hand graph shows testing on the cross validation data set - note the considerably worse performance of the delta rule.

learning is only affected by current neural activity. The average performance in this case was 42% on the training set and 35% on the cross validation set, highlighting the dramatic change in the classification performance of the competitive network when a short memory of previous activity is included.

2.3 Experiment II: Reconstructing the Receptive Fields of the Output Cells

In order to better understand the relative failure of the delta rule in the cross-validation phase of testing and conversely the relative success of the trace rule, the receptive fields of the output neurons are reconstructed. Reconstruction of the receptive fields of the output neurons is effected by projecting the weighted connections of each layer 2 neuron to each layer 1 neuron. These weightings are then multiplied by the according weights from the layer1 neurons to the input. By summing all of the projected connection weightings a reconstructed optimal input, or overall receptive field for the output neuron is generated. This process is limited by the effectiveness of linear reconstruction through a nonlinear system, but, the general process should be informative and at least indicative of the general form of the receptive fields of each output neuron.

2.4 Experiment II: Results

The results of this reconstruction procedure for three of the layer 2 neurons are shown in figure 4, the results of which are striking.

In the three examples shown, the central column represents the true mathematical average of the ten digits in each class, and shows remarkable accord with the receptive fields set up in the layer 2 neurons trained with the trace rule. This lends support to the idea that the trace rule is capable of training neurons to respond to the within-class average of a stimulus set. The results of the reconstruction process for the delta rule are also striking in that they show little or no inherent structure - supporting the earlier suggestion that weighting has been somewhat arbitrarily assigned to a few pixels in the input which happen to solve the classification problem for the training set - but which consequently result in poor generalization to novel digits[3].

2.5 Experiment I and II: Conclusions

The previous two experiments have sought to highlight and investigate the relative failure of a supervised learning rule to produce a robust solution to the digit classification problem posed. Under the constraints of the training and testing paradigm set out in these experiments, in which only a few training examples are provided, the cross validation performance of the delta rule has been seen to be significantly worse than that achieved by the trace rule.

Of course, it would be wrong to assume from these results that the delta rule represents the best performance of a weight specific tutored system. The delta rule, as proposed here, is in practice a somewhat naïve approach, and by saying this the intention is to imply that there are ways and means of preventing a tutored learning rule from over-fitting the data. Indeed, there are several principled mechanisms for improving the performance of nets trained with a delta rule signal, such as simply setting a training stopping point at a value below perfect performance, 95% say. However, such a procedure would not work in this case, since in practice, the performance on the cross validation set increases monotonically with performance on the training set - see figure 3.

In summary, the delta rule concentrates much of the neurons' powers of discrimination upon a few inputs which leads, ultimately, to classification errors on a novel set of stimuli. The trace rule however, produces a robust solution by learning something approaching the within-class average of each digit class. Because of the inherent structural similarity of the digits, the within-class average turns out to be a better strategy for learning about general digit shape.

[3]Note that synapses are allowed to take negative values under the delta rule scheme which means that dark areas on the reconstructed receptive fields represent regions of significant inhibitory input to the neuron.

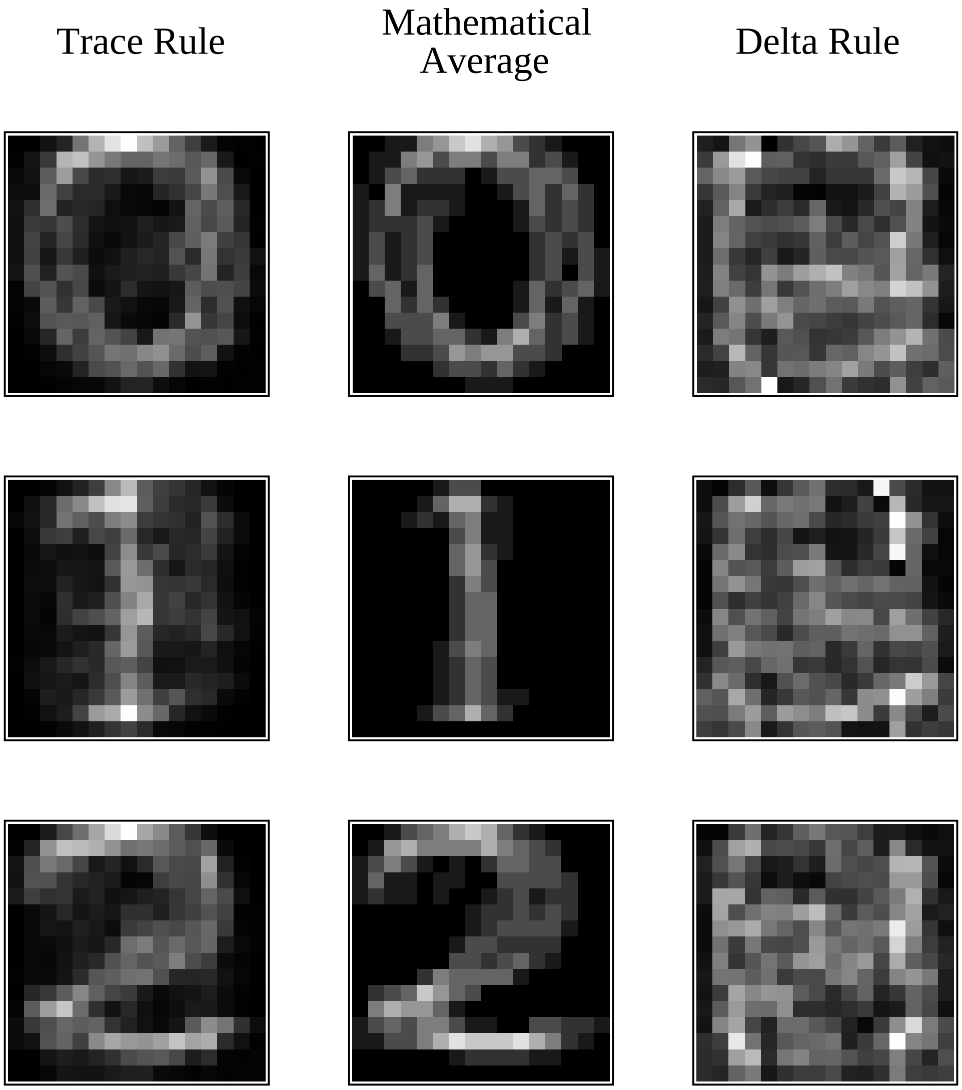

Figure 4: Reconstruction results for neurons responsive to digits 0, 1 and 2. The central column shows the true mathematical average of the ten test digits within each class.

3 Relative Performance of the Trace Rule and TraceNet Network Architecture

3.1 Experiment III: The Neocognitron and TraceNet

Attention now turns to measuring the relative performance of the TraceNet, trained with the trace rule in a digit classification task, as compared to the Neocognitron, a well established architecture originally proposed by Fukushima (1980; 1988; 1992).

In order to make a fair comparison between architectures, one would normally make reference to the following four performance ratings and network parameters as the basic criteria for comparison:

- Recognition performance of the network on the training set - i.e. how well the network does at recognizing digits it has seen before.
- Level of misclassification of digits by the network - important if a mechanism exists for rejecting stimuli about whose class the network is extremely 'unconfident'.
- Recognition performance of the network on a cross validation set - i.e. how well does the network perform on a set of digits drawn from the same data base, but which it has never seen before.
- Network size, both in terms of the number of units and the level of connectivity.

Lovell & Tsoi (1992) have carried out extensive testing of the Neocognitron and their data will form the basis of the comparison in performance of the two networks on the basis of first and last of the criteria listed above. The reasons for limiting the scope of the comparison are two-fold. Firstly, Lovell & Tsoi only relate net performance statistics for testing on a training set, preventing any discourse on relative cross validation performance. Secondly, the TraceNet simulation has no mechanism for rejecting unrecognized stimuli and therefore always 'recognizes' the stimulus present - albeit incorrectly! - which means that it is not possible to say much about the third criterion. Lovell & Tsoi's model does possess such a mechanism, but in practice the rejection rate for the most successful net is zero anyway, so mitigating any possible problems in forming a comparison on that point.

In order to find out how well the TraceNet performs on the identical 400 digit test set employed by Lovell & Tsoi, the complete set of the 400 digits were presented to the network 800 times - with the constraint that all examples of each digit class appeared in unbroken, though random sequences, as in the previous experiment. The digits used appear in figure 5. The new digit set is both larger and contains more starkly transformed exemplars and so despite the increased image dimensions- now 20x20 pixels - the set is not linearly separable. The dimensions of the network were

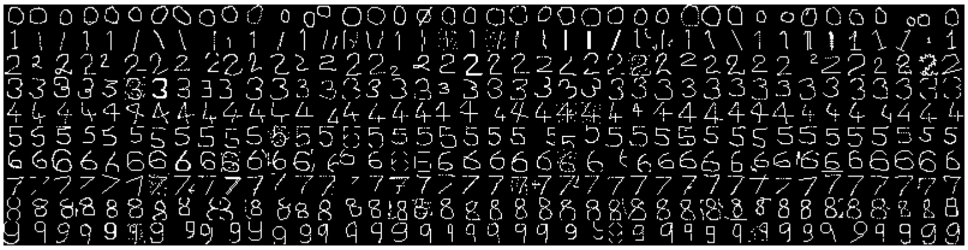

Figure 5: The training set used by Lovell & Tsoi (1992) in their work on the Neocognitron.

slightly enlarged to accommodate the larger characters appearing in the character set - with the input retina being increased to 20x20 pixels and the first layer from 16 4x4 pools to 25. The value of η was also changed (raised to 0.9), to reflect the increased number of examples in each stimulus class and hence the length of a single presentation sequence.

3.2 Experiment III: Results

A graph of recognition performance on each digit class appears in figure 6. Performance is now measured directly as the percentage correct assuming that each neuron classifies only one stimulus class - so as to make a direct comparison with the Lovell & Tsoi results possible. The best performance quoted by Lovell & Tsoi for their version of the Neocognitron, using Fukushima's *supervised* training algorithm (Fukushima, 1988), is 72.50%. The TraceNet's best performance using the trace rule is 74.5%. As a final comparison, the number of neurons in Fukushima's model is 11,320, with some 1.2 million connections. This compares with (4x4)x(5x5) layer 1 cells, and 5x2 output cells, in the TraceNet architecture - some 410 neurons in total, containing a total of 10,400 connections. Ofcourse, this figure is still relatively large when compared to the size of the training set - some 400 exemplars. However, the size is also a function of the dimensionality of the input space chosen originally by Lovell & Tsoi and which has been duly matched here, and the 10,400 connections still represent a 100 fold decrease in the number of weights used in the Neocognitron model. Although training times were not reported in the Lovell & Tsoi paper one would undoubtedly see a comparable saving in training time.

3.3 Experiment III: Conclusions

The slight improvement in performance of this simple network over the Neocognitron is too small to be statistically significant, however, in terms of network size and hence training time the trace net is shown to yield considerable savings over that of the Neocognitron. Of course, the results quoted by LeCun *et. al.*, of some

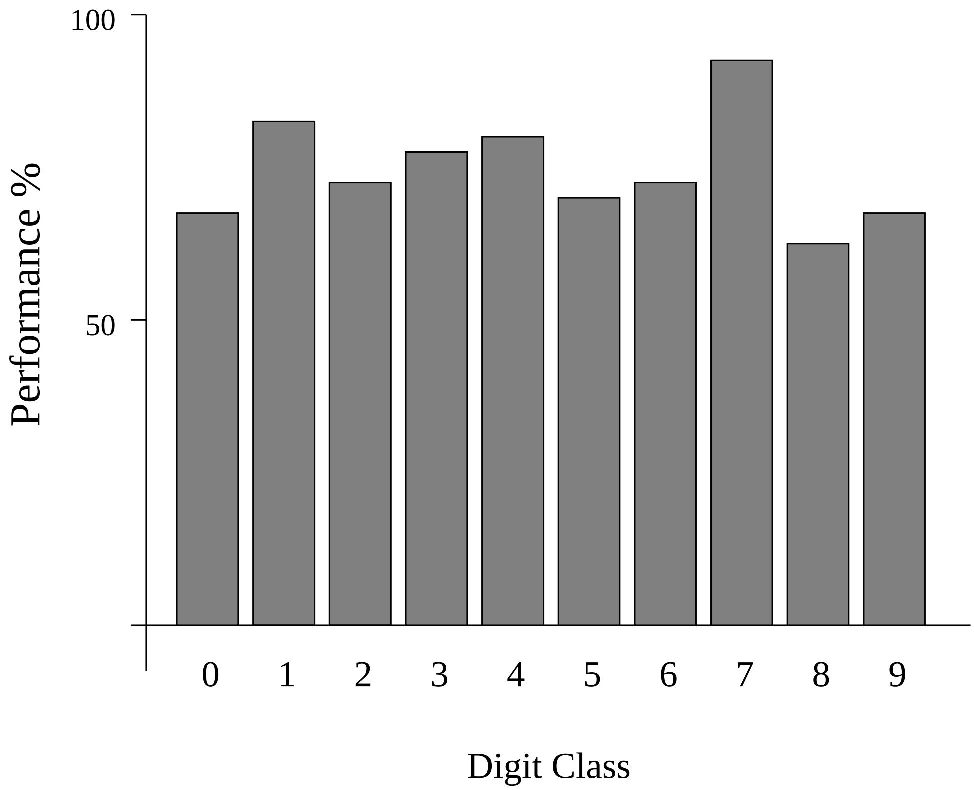

Figure 6: Recognition performance of the trace net for each digit class from the digit set used originally by Lovell & Tsoi (1992).

0.14% error on a training set of over 7000 digits, puts both results into perspective. But in fairness, their network is trained in a wholly supervised fashion, and their network utilizes some 1256 neurons, with over 66,000 connections, and in addition the stimuli used by LeCun *et. al.* are considerably better conditioned than those used by Lovell & Tsoi.

4 Discussion

The trace rule has been shown to outperform both the delta rule and the supervised version of the Neocognitron on a pair of handwritten character recognition tasks, its success being of particular interest because of its simple unsupervised form. Indeed, the possible role of such a learning rule in real neurons in inferior temporal (IT) cortex has been the focus of work reported elsewhere (Földiák, 1992; Wallis, 1994; Wallis & Rolls, 1995). IT cortex incorporates several visual processing areas strongly implicated in object recognition in primates (Goodale & Milner, 1992; Rolls, 1992; Tanaka *et al.*, 1991).

Bearing this biological motivation of the trace rule in mind, it is clear that it would ordinarily be used in a rather different context from that of character recognition. It is best suited to solving a classification task in which one has no external signal to associate any particular image with any particular object class, but in which sequences of images belonging to a single object are viewed in sequences - such as we experience with everyday objects all of our lives. Evidence that we might categorise images based upon the temporal statistics of our environment has received some support for the work by Miyashita (1988). He illustrated that stimuli can be arbitrarily associated together by a single IT neuron, distinct from many other spatially similar images simply due to their association not in space, but in time.

The learning of digit categories may, in fact, be a special case in that it may be largely a supervised process for humans - starting with a real human teacher. However, by training the network on this task it has been possible to favorably compare the classification performance of this learning rule and simple network architecture with other well established rules and architectures, and further, to illustrate how a simple modification to the basic Hebbian learning scheme can drastically alter the types of categorisation which can be achieved.

References

Attneave, F. 1959. *Applications of Information Theory to Psychophysics: A Summary of Basic Concepts, Methods and Rsults.* New York: Holt, Rinehart and Winston.

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

Földiák, P. 1992. *Models of Sensory Coding.* Tech. rept. CUED/F–INFENG/TR 91. University of Cambridge, Department of Engineering.

Fukushima, K. 1980. Neocognitron: A self–organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, **36**, 193–202.

Fukushima, K. 1988. Neocognitron: A hierarchical neural network model capable of visual pattern recognition unaffected by shift in position. *Neural Networks*, **1**, 119–130.

Fukushima, K. 1992. Character recognition with neural networks. *Neurocomputing*, **4**(5), 221–233.

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

Hecht-Nielsen, R. 1990. *Neurocomputing.* Reading, Massachusetts: Addison-Wesley.

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

Klopf, A.H. 1988. A neuronal model of classical conditioning. *Psychobiology*, **16**, 85–125.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., & Jackel, L.D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, **1**, 541–551.

Lovell, D.R., & Tsoi, A.C. 1992. *The performance of the Neocognitron with various S-cell and C-cell transfer functions.* Paper posted to the 'Connectionists' internet mail group. Produced in the Dept. Electrical Engineering, University of Queensland, Queensland 4072, Australia.

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

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

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. 1986. Learning internal representations by error propagation. *Chap. 8 of:* Rumelhart, D.E., & McClelland, J.L. (eds), *Parallel Distributed Processing*, vol. 1: Foundations. Cambridge, Massachusetts: MIT Press.

Sutton, R.S., & Barto, A.G. 1981. Towards a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, **88**, 135–170.

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

Wallis, G. 1994. *Neural Mechanisms Underlying Processing in the Visual Areas of the Occipital and Temporal Lobes.* Ph.D. thesis, Department of Experimental Psychology, Oxford Univeristy. www: ftp://ftp.mpik-tueb.mpg.de/pub/guy/all.ps.Z.

Wallis, G., & Rolls, E.T. 1995. A Model of Invariant Object Recognition in The Visual System. *Submitted to Journal of Computational Neuroscience for Review.* www: ftp://ftp.mpik-tueb.mpg.de/pub/guy/jcns7.ps.Z.

Wallis, G., Rolls, E.T., & Földiák, P. 1993. Learning invariant responses to the natural transformations of objects. *Pages 1087–1090 of: International Joint Conference on Neural Networks*, vol. 2.

Widrow, B., & Hoff, M.E. 1960. Adaptive switching circuits. *Pages 96–104 of: 1960 IRE WESCON Convention Record*, vol. 4. New York: Institute of Radio Engineers.