To appear in H. Wechsler, P.J. Phillips, V. Bruce, S. Fogelman-Soulie, T. Huang (Eds.), Face Recognition: From Theory to Applications, NATO ASI Series F. Springer-Verlag.

Learning Viewpoint Invariant Face Representations from Visual Experience by Temporal Association

Marian Stewart Bartlett[1] and Terrence J. Sejnowski[2]

[1]Departments of Cognitive Science and Psychology, UCSD, San Diego, CA 92093 and The Salk Institute, La Jolla, CA 92037. marni@salk.edu
[2]Department of Biology, UCSD, San Diego, CA 92093, and Howard Hughes Medical Institute at The Salk Institute, La Jolla, CA 92037. terry@salk.edu

Abstract. In natural visual experience, different views of an object or face tend to appear in close temporal proximity. A set of simulations is presented which demonstrate how viewpoint invariant representations of faces can be developed from visual experience by capturing the temporal relationships among the input patterns. The simulations explored the interaction of temporal smoothing of activity signals with Hebbian learning (Foldiak, 1991) in both a feed-forward system and a recurrent system. The recurrent system was a generalization of a Hopfield network with a lowpass temporal filter on all unit activities. Following training on sequences of graylevel images of faces as they changed pose, multiple views of a given face fell into the same basin of attraction, and the system acquired representations of faces that were approximately viewpoint invariant.

1 Introduction

Cells in the primate inferior temporal lobe have been reported that respond selectively to faces despite substantial changes in viewpoint (Perrett, Mistlin, & Chitty, 1989; Hasselmo, Rolls, Baylis, & Nalwa, 1989). A small proportion of cells gave responses that were invariant to angle of view, whereas other cells had tuning curves that were quite broad. Perrett et al. (1989) reported broad coding for five principal views of the head: Frontal, left profile, right profile, looking up, and looking down, and the pose tuning of these cells was on the order of $\pm 40^0$. The retinal input changes considerably under these shifts in viewpoint.

This model addresses how receptive fields with such broad pose tuning could be developed from visual experience. The model touches on several issues in the psychology and neurophysiology of face recognition. Can general learning principles account for the ability to respond to faces across changes in pose, or does this function require special purpose, possibly genetically encoded mechanisms? Is it possible to recognize faces across changes in pose without explicitly recovering or storing the 3-dimensional structure of the face? What are the potential contributions of temporal sequence information to the representation and recognition of faces?

Until recently, most investigations of face recognition focused on static images of faces. The preponderance of our experience with faces, however, is not with static faces, but with live faces that move, change expression, and pose. Temporal sequences contain information that can aid in the process of representing and recognizing faces and objects (eg. see V. Bruce, this volume). This model explores how a neural system can acquire invariance to viewpoint from visual experience by accessing the temporal structure of the input. The appearance of an object or a face changes continuously as the observer moves through the environment or as a face changes expression or pose. Capturing the temporal relationships in the input is a way to automatically associate different views of an object without requiring three-dimensional representations (Stryker, 1991).

Temporal association may be an important factor in the development of transformation invariant responses in the inferior temporal lobe of primates (Rolls, 1995). Neurons in the anterior inferior temporal lobe are capable of associating patterns by temporal proximity. After prolonged exposure to a sequence of randomly generated fractal patterns, correlations emerged in the sustained responses to neighboring patterns in the sequence (Miyashita, 1988). Macaques were presented a fixed sequence of 97 fractal patterns for 2 weeks. After training, the patterns were presented in random order. Figure 1 shows correlations in sustained responses of the AIT cells to pairs of patterns as a function of the relative position of the patterns in the training sequence. Responses to neighboring patterns were correlated, and the correlation dropped off as the distance between the patterns in the training sequence increased. These data suggest that cells in the temporal lobe can modify their receptive fields to associate patterns that occurred close together in time.

Temporal relationships can be captured by Hebbian learning (Foldiak, 1991). Hebbian learning is an unsupervised learning rule that was proposed as a model for the modification of synaptic strengths between neurons (Hebb, 1949). In Hebbian learning, the connections between simultaneously active units are strengthened. With a lowpass temporal filter on output unit activities, Hebbian learning will strengthen the connections between active inputs and *recently* active outputs. This mechanism can learn transformation invariant representations when different views of an object are presented in temporal continuity (Foldiak, 1991; Weinshall, Edelman & Bulthoff, 1991; Rhodes, 1992; O'Reilly & Johnson, 1994). Such learning mechanisms have recently been shown to learn transformation invariant of responses to complex inputs such as images of faces (Bartlett & Sejnowski, 1996, 1997; Wallis & Rolls, 1997; Becker, 1997).

There are several mechanisms by which receptive fields could be modified to perform temporal associations. The NMDA channel is an activity dependent gate in the neural membrane thought to be involved in long-term potentiation of synaptic strengths. A temporal window for Hebbian learning could be provided by the 0.5 second open-time of the NMDA channel after the neuron fires (Rhodes, 1992; Rolls, 1992). Reciprocal connections between cortical regions (O'Reilly & Johnson, 1994) or lateral interconnections within cortical regions could sustain activity over longer time periods and allow temporal associations across larger time scales.

The time course of the modifiable state of a neuron, based on the open time of the NMDA channel, has been modeled by a lowpass temporal filter on the post-synaptic unit activities (Rhodes, 1992). This paper examines the contribution of such temporal smoothing to the development of viewpoint invariant responses in both a feedfor-

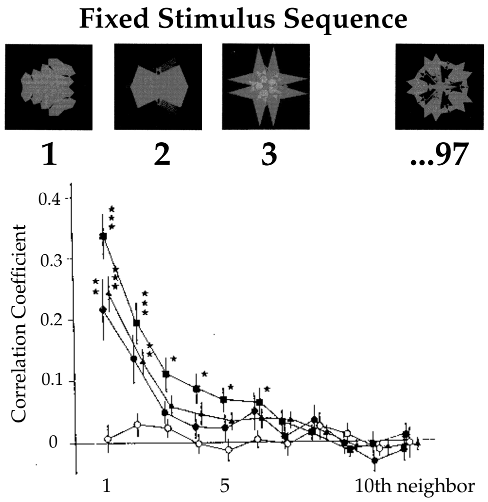

Figure 1: Evidence of temporal associations in IT. Top: Samples of the 97 fractal pattern stimuli in the fixed training sequence. Stimuli were in color. Bottom: Autocorrelograms on the sustained firing rates of AIT cells along the serial position number of the stimuli. Abscissa is the relative position of the patterns in the training sequence. Triangles are mean correlations in responses to the learned stimuli for 57 cells. Open circles are correlations in responses to novel stimuli for 17 cells, and closed circles are responses to learned stimuli for the same 17 cells. Squares are mean correlations for the 28 cells with statistically significant response correlations, according to Kendall's correlation test. Adapted from Miyashita (1988). Reprinted with permission from *Nature*, copyright 1988, MacMillan Magazines Ltd.

ward and a recurrent system. In the feedforward system, the Competitive Learning rule (Rumelhart & Zipser, 1985) is extended to incorporate an activity trace on the output unit activities (Foldiak, 1991). The recurrent component of the simulation examines the development of temporal associations in an attractor network. Perceptual representations have been related to to basins of attraction in activity patterns across an assembly of cells (Amit, 1995; Freeman, 1994; Hinton & Shallice, 1991). The simulation performed here shows how viewpoint invariance can be captured in an attractor network representation. The recurrent system was a generalization of a Hopfield network with a lowpass temporal filter on all unit activities (Hopfield, 1982). We show that the combination of basic Hebbian learning with temporal smoothing of unit activities produces a generalization of an attractor network learning rule that associates temporally proximal input patterns into basins of attraction (Griniasty, Tsodyks, & Amit, 1993). Following training on sequences of graylevel images of faces as they changed pose, multiple views of a given face fell into the same basin of attraction, and the system acquired representations of faces that were approximately viewpoint invariant.

2 Simulation

Stimuli for these simulations consisted of 100 images of faces undergoing a change in pose, from Beymer (1994). There were twenty individuals at each of five poses, ranging from $-30°$ to $30°$. The faces were automatically located in the frontal view image by using a feature-based template matching algorithm (Beymer, 1994). The images were normalized for luminance and scaled to 120 x 120.

Images were presented to the model in sequential order as the subject changed pose from left to right (Figure 2). The first layer of processing consisted of an oriented energy model related to the output of V1 complex cells (Daugman, 1988; Lades et al., 1993; Heeger, 1991). The images were filtered by a set of sine and cosine Gabor filters at 4 spatial scales (32, 16, 8, and 4 pixels per cycle), and at four orientations (vertical, horizontal, and $\pm 45°$.) The outputs of the sine and cosine Gabor filters were squared and summed, and the result was sampled at 8 pixel intervals.

The set of V1 model outputs projected to a second layer of 70 units labeled "complex pattern units" to characterize their receptive fields after learning. The 70 units were grouped into two pools, and there was feedforward inhibition between all of the units in a pool. The complex pattern unit activities were passed through a lowpass temporal filter, described below. The third stage of the model was an attractor network produced by lateral interconnections among all of the complex pattern units. The feedforward and lateral connections were updated successively.

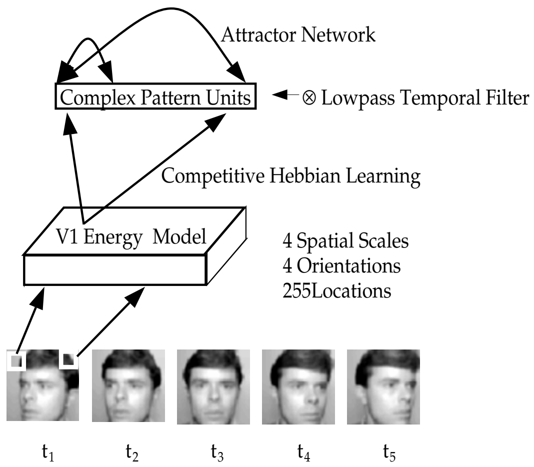

Figure 2: Model architecture.

2.1 Competitive Hebbian learning of temporal relationships

The learning rule for the feedforward connections of the model was an extension of the Competitive Learning Algorithm (Rumelhart & Zipser, 1985; Grossberg, 1976) in which the output unit activities were passed through a lowpass temporal filter. This manipulation gave the winner in the previous time steps a competitive advantage for winning, and therefore learning, in the current time step.

Let y_j^t be the weighted sum of the feedforward inputs. The output activity of unit j at time t, $\overline{y_j}^{(t)}$, is determined by the trace, or running average, of its input activity:

$$\overline{y_j}^{(t)} = (1-\lambda)y_j^t + \lambda\overline{y_j}^{(t-1)} \tag{1}$$

The output unit activity, V_j, was subject to a step-nonlinear competition function.

$$V_j = \begin{cases} 1 & \text{if } j = max_j[\overline{y_j}^{(t)}] \\ 0.1 & \text{otherwise} \end{cases} \tag{2}$$

The connections were updated according to the following learning rule:

$$\Delta w_{ij} \propto V_j\left(\frac{x_{iu}}{\sum_k x_{ku}} - w_{ij}\right) \tag{3}$$

The weight change from input i to output j was proportional to x_{iu}, the input activity at unit i for pattern u, normalized by the total input activation for pattern u, minus a weight decay term. The weight to each unit was constrained to sum to one. Note that the output unit with the most activity for the current input pattern learned an order of magnitude more than the other units. One face image was input to the system per time step. The temporal smoothing was subject to a line process. There was no temporal smoothing across time steps in which the magnitude of the difference between subsequent input images was large.

The competitive learning rule alone, without the temporal smoothing, partitioned the set of inputs into roughly equal groups by spatial similarity. With the temporal smoothing, this learning rule clustered the input by a combination of spatial similarity and temporal proximity, where the relative contribution of the two factors was determined by the parameter λ. This learning rule is related to spatio-temporal principal component analysis. Competitive Hebbian learning can find the principal components of the input data (Oja, 1989; Sanger, 1989). The low-pass temporal filter on output unit activities in Equation 1 causes Hebbian learning to find axes along which the data covaries over recent temporal history.

2.2 Temporal association in an attractor network

The lateral interconnections in the output layer formed an attractor network. After the feedforward connections were established, the weights of the lateral connections were trained with a basic Hebbian learning rule. Hebbian learning of lateral interconnections, in combination with the lowpass temporal filter on the unit activities in (1), produced a learning rule that associated temporally proximal inputs into basins of attraction.

This is demonstrated as follows. We begin with a basic Hebbian learning rule:

$$W_{ij} = \frac{1}{N}\sum_{t=1}^{P}(y_i^t - y^0)(y_j^t - y^0) \tag{4}$$

where N is the number of units, P is the number of patterns, and y^0 is a baseline activity rate which we define as the mean activity over all of the units. Replacing y_i^t with the activity trace $\overline{y_i}^{(t)}$ defined in Equation 1, substituting $y^0 = \lambda y^0 + (1-\lambda)y^0$ and multiplying out the terms produces the following learning rule:

$$W_{ij} = \frac{1}{N}\sum_{t=1}^{P}(1-\lambda)^2(y_i^t - y^0)(y_j^t - y^0)$$

$$+\lambda(1-\lambda)\left[(y_i^t - y^0)(\overline{y_j}^{(t-1)} - y^0) + (\overline{y_i}^{(t-1)} - y^0)(y_j^t - y^0)\right]$$

$$+\lambda^2\left[(\overline{y_i}^{(t-1)} - y^0)(\overline{y_j}^{(t-1)} - y^0)\right] \tag{5}$$

This learning rule is a generalization of an attractor network learning rule that has been shown to associate random input patterns into basins of attraction based on serial position in the input sequence (Griniasty, Tsodyks & Amit, 1993). The first term in this equation is basic Hebbian learning. The weights are proportional to the covariance matrix of the input patterns at time t. The second term performs Hebbian association between the patterns at time t and $t-1$. The third term is Hebbian association of the trace activity for pattern $t-1$.

The following update rule was used for the activation V of unit i at time t from the lateral inputs (Griniasty, Tsodyks, & Amit, 1993):

$$V_i(t+\delta t) = \phi\left[\sum W_{ij}V_j(t) - \theta\right] \tag{6}$$

Where θ is a neural threshold and $\phi(x) = 1$ for $x > 0$, and 0 otherwise. In these simulations, $\theta = 0.007$, $N = 70$, $P = 100$, $y^0 = 0.03$, and $\lambda = 0.5$.

2.3 Results

The feedforward and the lateral connections were trained successively. Sequences of face images were presented to the network in order as each subject changed pose from left to right. The feedforward connections were updated by the learning rule in Equation 3, with $\lambda = 0.5$. Competitive interactions were among two pools of 35 units, such that two units were active for any given face.

After training the feedforward connections, the representation of each face was a sparse representation consisting of the two active output units of the 70 complex pattern units. Network output was evaluated with the temporal filter removed. "Pose tuning" of the feedforward system was assessed by comparing correlations in the population activities for different views of the same face to correlations across faces of different people (Figure 3 Left). Pose tuning is shown both with and without the temporal lowpass filter on unit activities during training. The temporal filter broadened the pose tuning of the feedforward system, producing a response that was more selective for the individual and less dependent on viewpoint.

The feedforward system trained with $\lambda = 0.5$ provided a sparse input to the attractor network. After the feedforward connections were established, the feedforward weights were held fixed, and sequences of face images were again presented to the network as each subject gradually changed pose. The lateral connections among the

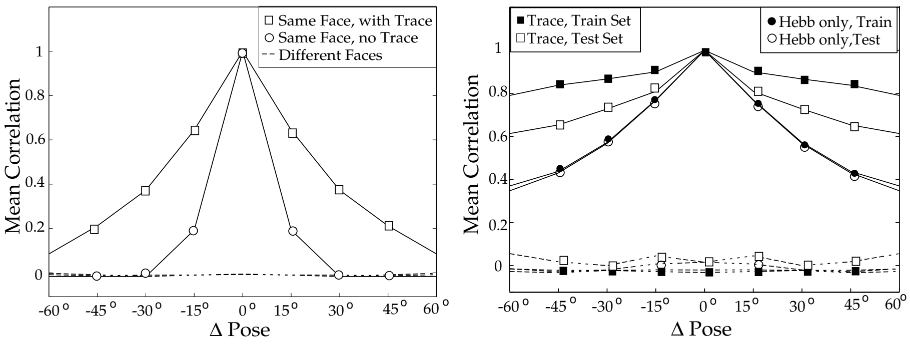

Figure 3: Left: Correlation of the outputs of the feedforward system as a function of change in pose. Correlations across different views of the same face (–) are compared to correlations across different faces (——) with the temporal trace parameter $\lambda = 0.5$ and $\lambda = 0$. Right: Correlations in sustained activity patterns in the attractor network as a function of change in pose. Results obtained with Equation 5 (Hebb plus trace) are compared to Hebb only. Closed symbols are training set results and open symbols are test set results.

output units were updated by the learning rule in Equation 5. After training the attractor network, each face was presented to the system, and the activities in the output layer were updated until they arrived at a stable state. Following learning, the patterns of sustained activity in the attractor network were approximately viewpoint invariant.

Figure 3 (Right) shows the correlations in the sustained activity patterns as a function of change in pose. The graph compares correlations obtained with Equation 5, using $\lambda = 0.5$, to that obtained with $\lambda = 0$, which is straight Hebbian learning. Test image performance was evaluated by alternately training on four poses and testing on the fifth, and then averaging the five test performances.

2.4 Weight structure and fixed points of the attractor network

The weight structure and fixed points of the attractor network are illustrated in Figure 4 using an idealized data set in order to facilitate visualization. The idealized data set contained 25 input patterns, where each pattern was coded by activity in a single bit (Figure 4, Top). The patterns represent 5 individuals with 5 views each. The middle graph in Figure 4 shows the weight matrix obtained with the attractor network learning rule, with $\lambda = 0.5$. Note the approximately square structure of the weights along the diagonal, showing positive weights among most of the 5 views of each individual. The inset shows the actual weights between views of individuals 3 and 4. The weights decreased with the distance between the patterns in the input sequence. The bottom graphs show the fixed points attained for each input pattern. Note that the same sustained pattern of activity was obtained no matter which of the 5 views of an individual was input to the network. For this idealized representation, the attractor network produced responses that were entirely viewpoint invariant.

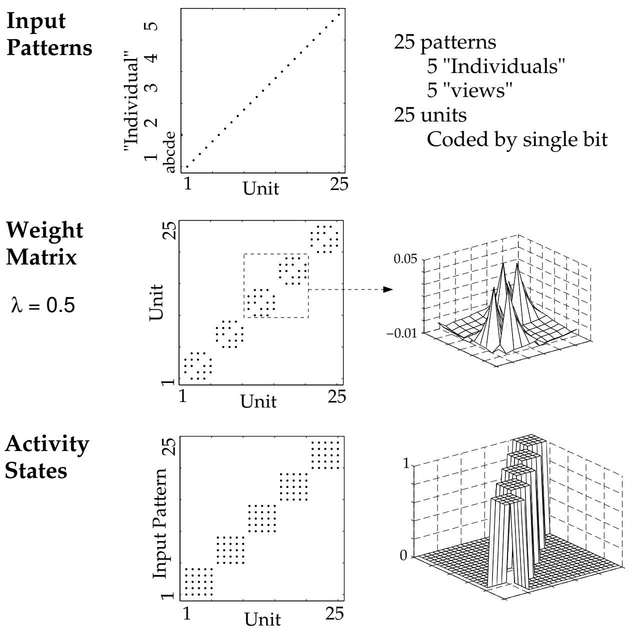

Figure 4: Demonstration of attractor network with idealized data. Top: Idealized data set. The patterns consist of 5 ”individuals” (1,2,3,4,5) with five ”views” each (a,b,c,d,e), and are each coded by activity in 1 of the 25 units. Center: The weight matrix obtained with equation 3. Dots show the locations of positive weights, and the inset shows the actual weights among the 5 views of two different individuals. Bottom: Fixed points for each input pattern. Unit activities are plotted for each of the 25 input patterns.

3 Discussion

Many cells in the primate anterior inferior temporal lobe and superior temporal sulcus maintain their response preferences to faces or three dimensional objects over substantial changes in viewpoint (Hasselmo, Rolls, Baylis, & Nalwa, 1989; Perrett Mistlin & Chitty, 1989; Logothetis & Pauls, 1995). This set of simulations demonstrated how such viewpoint invariant representations could be developed from visual experience through unsupervised learning.

This simulation began with structured inputs similar to the responses of V1 complex cells, and explored the performance of unsupervised learning mechanisms that can transform these inputs into pose invariant responses. We showed that a lowpass temporal filter on unit activities, which has been related to the time course of the modifiable state of a neuron (Rhodes, 1992), cooperates with Hebbian learning to (1) increase the viewpoint invariance of responses to faces in a feedforward system, and (2) create basins of attraction in an attractor network which associate temporally proximal inputs. This simulation demonstrated how viewpoint invariant representations of complex objects such as faces can be developed from visual experience by accessing the temporal structure of the input. The model addressed potential roles for both feedforward and lateral interactions in the self-organization of object representations, and demonstrated how viewpoint invariant responses can be learned in an attractor network.

Temporal sequences contain information that can aid in the process of representing and recognizing faces and objects. This model presented a means by which temporal information can be incorporated in the representation of a face. The learning rule in the feedforward component of this model extracted information about how the Gabor filter outputs covaried in recent temporal history in addition to how they covaried over static views. The processing in this model was related to spatio-temporal principal component analysis of the Gabor filter representation.

In this model, pose invariant face recognition was acquired by learning associations between 2-dimensional patterns, without recovering 3-D coordinates or structural descriptions. It has been proposed that 3-D object recognition may be performed through exposure to multiple 2-dimensional views and may not require the formation of internal 3-dimensional models, as was previously assumed (Poggio & Edelman, 1990; Ullman & Basri, 1991; Bulthoff, Edelman & Tarr, 1995). Such view-based representations may be particularly relevant for face processing, given the psychophysical evidence presented in this volume for face representations based on low-level filter outputs (see Biederman, Bruce, this volume).

In example-based models of recognition such as radial basis functions (Poggio & Edelman, 1990), neurons with view-independent responses are proposed to pool responses from view-dependent neurons. This model suggests a mechanisms for how this pooling could be learned. Logothetis and Pauls (1995) reported a small percentage of viewpoint invariant responses in the AIT of monkeys that were trained to recognize wire-framed objects across changes in view. The training images in this study oscillated $\pm 10^0$ from the vertical axis. The temporal association hypothesis presented in this paper suggests that more viewpoint invariant responses would be recorded if the monkeys were exposed to full rotations of the objects during training.

Acknowledgments

This project was supported by Lawrence Livermore National Laboratory ISCR Agreement B291528, and by the McDonnell-Pew Center for Cognitive Neuroscience at San Diego.

References

Amit, D. (1995). The Hebbian paradigm reintegrated: Local reverberations as internal representations. *Behavioral and Brain Sciences* 18:617-657.

Bartlett, M. Stewart, & Sejnowski, T., 1996. Unsupervised learning of invariant representations of faces through temporal association. *Computational Neuroscience: Int. Rev. Neurobio. Suppl. 1* J.M Bower, Ed., Academic Press, San Diego, CA:317-322.

Bartlett, M. Stewart, & Sejnowski, T., 1997. Viewpoint invariant face recognition using independent component analysis and attractor networks. In M. Mozer, M. Jordan, & T. Petsche, Eds., *Advances in Neural Information Processing Systems 9.* Cambridge, MA: MIT Press, p. 817-823.

Becker, S. (1997). Learning temporally persistent hierarchical representations. In M. Mozer, M. Jordan, & T. Petsche, Eds., *Advances in Neural Information Processing Systems 9.* Cambridge, MA: MIT Press, p. 824-830.

Beymer, D. 1994. Face recognition under varying pose. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition.* Los Alamitos, CA: IEEE Comput. Soc. Press: 756-61.

Biederman, I. (in press). Neural and psychophysical analysis of object and face recognition. In H . Wechsler and V. Bruce, Eds., *Face Recognition: From Theory to Applications.* Springer-Verlag.

Bulthoff, H.H. Edelman, S.Y., & M.J. Tarr (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex* 3 247-260.

Bruce, V. (in press). Human face perception and identification. In H . Wechsler and V. Bruce, Eds., *Face Recognition: From Theory to Applications.* Springer-Verlag.

Daugman, J.G. (1988). Complete discrete 2D Gabor transform by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 36: 1169-1179.

Foldiak, P. 1991. Learning invariance from transformation sequences. *Neural Computation* 3:194-200.

Freeman, W.J. 1994. Characterization of state transitions in spatially distributed, chaotic, nonlinear, dynamical systems in cerebral cortex. *Integrative Physiological and Behavioral Science*, 1994 29(3):294-306.

Griniasty, M., Tsodyks, M., & Amit, D. 1993. Conversion of temporal correlations between stimuli to spatial correlations between attractors. *Neural Comp.* 5:1-17.

Hasselmo M. Rolls E. Baylis G. & Nalwa V. 1989. Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research* 75(2):417-29.

Hebb, D. (1949). *The organization of behavior*. New York: Wiley.

Heeger, D. (1991). Nonlinear model of neural responses in cat visual cortex. *Computational Models of Visual Processing*, M. Landy & J. Movshon, Eds. MIT Press, Cambridge, MA.

Hinton, G. & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review* 98(1):74-75.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA* 79 2554-2558.

Lades, M., Vorbruggen, J., Buhmann, J., Lange, J., Konen, W., von der Malsburg, C., and Wurtz, R. (1993): Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers* 42(3): p. 300-311.

Oja, E. (1989). Neural networks, principal components, and subspaces. *International Journal of Neural Systems* 1(1): 61-68.

Perrett, D. Mistlin, A. & Chitty, A. (1989). Visual neurones responsive to faces. *Trends in Neuroscience* 10: 358-364.

Poggio, T. & Edelman, S. (1990). A network that learns to recognize 3-dimensional objects. *Nature* 343: 263-266.

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* 335(27); p.817-820.

Logothetis, N. & Pauls 1995. *Cerebral Cortex* Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cerebral Cortex* 3: 270-288.

O'Reilly, R. & Johnson, M. 1994. Object recognition and sensitive periods: A computational analysis of visual imprinting. *Neural Computation* 6:357-389.

Rhodes, P. 1992. The long open time of the NMDA channel facilitates the self-organization of invariant object responses in cortex. *Soc. Neurosci. Abst.* 18:740.

Rolls, E.T. (1995). Learning mechanisms in the temporal lobe visual cortex. *Behav. Brain Research* 66; p. 177-185.

Rumelhart, D. & Zipser, D. 1985. Feature discovery by competitive learning. *Cognitive Science* 9: 75-112.

Sanger, T. (1989a). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks* 2, 459-473.

Stryker, M. 1991. Temporal Associations. *Nature* 354:108-109.

Tsodyks M. & Feigel'man M. (1988). The enhanced storage capacity in neural networks with low activity level. *Europhysics Letters* 101-105.

Wallis, G; Rolls, E T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology (Oxford)*, 51(2): 167-194.

Weinshall, D.& Edelman, S. 1991. A self-organizing multiple view representation of 3D objects. *Bio. Cyber.* 64(3):209-219.