Learning viewpoint invariant face representations from visual experience in an attractor network

Marian Stewart Bartlett[†] and Terrence J. Sejnowski[‡]

[†]University of California San Diego Departments of Cognitive Science and Psychology; The Salk Institute, 10010 North Torrey Pines Road, La Jolla, CA 92037. Email: marni@salk.edu.

[‡]University of California San Diego Department of Biology; Howard Hughes Medical Institute at the Salk Institute, 10010 North Torrey Pines Road, La Jolla, CA 92037. Email: terry@salk.edu

Short title: Learning viewpoint invariant face representations
In press: *Network: Computation in Neural Systems* 9(3), 399-417, 1998.

Abstract

In natural visual experience, different views of an object or face tend to appear in close temporal proximity as an animal manipulates the object or navigates around it, or as a face changes expression or pose. 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 (Földiák, 1991) in both a feedforward layer and a second, recurrent layer of a network. The feedforward connections were trained by Competitive Hebbian Learning with temporal smoothing of the post-synaptic unit activities (Bartlett & Sejnowski, 1996). The recurrent layer was a generalization of a Hopfield network with a lowpass temporal filter on all unit activities. The combination of basic Hebbian learning with temporal smoothing of unit activities produced an attractor network learning rule that associated temporally proximal input patterns into basins of attraction. These two mechanisms were demonstrated in a model that took graylevel images of faces as input. Following training on image sequences 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 that have been classed as viewpoint *dependent* 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 (e.g. Bruce, in press). 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 pose invariant responses in the inferior temporal lobe of primates (Rolls, 1995). Neurons in the anterior inferior temporal lobe are capable of forming temporal associations in their sustained activity patterns. 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.

Hebbian learning can capture temporal relationships in a feedforward system when the output unit activities undergo temporal smoothing (Földiák, 1991). This mechanism learns viewpoint-tolerant representations when different views of an object are presented in temporal continuity (Földiák, 1991; Weinshall & Edelman, 1991; Rhodes, 1992; O'Reilly & Johnson, 1994; Wallis & Rolls, 1997). Földiák (1991) used temporal association to model the development of viewpoint invariant responses of complex V1 cells from sweeps of oriented edges across the retina. This model achieved translation invariance in a single layer by having orientation-tuned filters in the first layer that produced linearly separable patterns. More generally, approximate viewpoint invariance may be achieved by the superposition of several Földiák-like networks (Rolls, 1995). Most such models used idealized input representations. These 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, in press). The assumption of temporal coherence can also be applied to learn other properties of the visual environment, such as depth from stereo disparity of curved surfaces (Becker, 1993; Stone, 1996).

There are several mechanisms by which receptive fields could be modified to perform temporal associations. A temporal window for Hebbian learning could be provided by the 0.5 second open-time of the NMDA channel (Rhodes, 1992; Rolls, 1992). A spatio-temporal window for Hebbian learning could also be produced by the release of a chemical signal following activity such as nitric oxide (Montague, Gally, & Edelman, 1991). Recurrent excitatory connections within a cortical area and reciprocal connections between cortical regions (O'Reilly & Johnson, 1994) 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 for calcium influx, has been modeled by a lowpass temporal filter on the post-synaptic unit activities (Rhodes, 1992). A lowpass temporal filter is a simple way to describe mathematically any of the above effects. This paper examines the contribution of such a lowpass temporal filter to the development of viewpoint invariant responses in both a feedforward layer, and a second, recurrent layer of a network. In the feedforward system, the Competitive Learning rule (Rumelhart & Zipser, 1985) is extended to incorporate an activity trace on the output unit activities (Földiák, 1991). The activity trace causes recently active output units to have a competitive advantage for learning subsequent input patterns.

The recurrent component of the simulation examines the development of temporal associations in an attractor network. Perceptual representations have been related to basins of attraction in activity patterns across an assembly of cells (Amit, 1995; Freeman, 1994; Hinton & Shallice, 1991). Weinshall and Edelman (1991) modeled the development of viewpoint invariant representations of wire-framed objects by associating neighboring views into basins of attraction. The simulations performed here show how viewpoint invariant representations of face images can be captured in an attractor network, and we examine the effect of a lowpass temporal filter on the attractor network learning rule. The recurrent layer was a generalization of a Hopfield network (Hopfield, 1982) with a lowpass temporal filter on all unit activities. We show that the combination of basic Hebbian learning with temporal smoothing of unit activities produces an attractor network learning rule that associates temporally proximal input patterns into basins of attraction. This learning rule is a generalization of an attractor network learning rule that produced temporal associations between randomly generated input patterns (Griniasty, Tsodyks, & Amit, 1993).

These two mechanisms were implemented in a model with both feedforward and lateral connections. The input to the model consisted of the outputs of an array of Gabor filters. These were projected through feedforward connections to a second layer of units, where unit activities are passed through a lowpass temporal filter. The feedforward connections were modified by competitive Hebbian learning to cluster the inputs based on a combination of spatial similarity and temporal proximity. Lateral connections in the output layer created an attractor network that formed basins of attraction based on the temporal proximity of the input patterns. 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) (see Figure 2). 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 location of the face in the frontal view image defined a window for the other images in the sequence. Each input sequence therefore consisted of a single stationary window within which the subject moved his or her head. The images were normalized for luminance and scaled to 120 x 120 pixels.

2.1 Model architecture

Images were presented to the model in sequential order as the subject changed pose from left to right (Figure 3). The first layer of processing consisted of an oriented energy model related to the output of V1 complex cells (Daugman, 1988; Lades, Vorbrüggen, Buhmann, Lange, Konen, von der Malsburg, and Würtz, 1993). 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 standard deviation of the Gaussian was set to twice the frequency of the sine or cosine wave, such that the receptive field size of the spatial filters increased with the spatial scale of the filters. The outputs of the sine and cosine Gabor filters were squared and summed, and then normalized by scale and orientation (Heeger, 1991). The result was sampled at 8 pixel intervals. This produced a 3600-dimensional representation consisting of 225 spatial locations, 4 spatial scales, and 4 orientations.

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 complex pattern unit activities were passed through a lowpass temporal filter, described below. There was feedforward inhibition between the complex pattern units, meaning that the competition influenced the feedforward activations only. The 70 units were grouped into two inhibitory pools, such that there were two active complex pattern units for any given input pattern. 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.

2.2 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). The output unit activities were passed through a lowpass temporal filter (Bartlett & Sejnowski, 1996). This manipulation gave active units in the previous time steps a competitive advantage for winning, and therefore learning, in the current time step.

Let $y_j^t = \sum_i w_{ij} x_i + b_j$ be the weighted sum of the feedforward inputs and the bias at time t. The 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.

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

where α is the learning rate, and N is the number of clustering units in the output layer. This was a modified winner-take-all competition where the non-winning activation was set to a constant small value rather than zero. The effect of the small positive activation was to cause non-winning weight vectors to move into the space spanned by the input data (Rumelhart & Zipser, 1985). The feedforward connections were updated according to the following learning rule:

$$\Delta w_{ij} = \alpha V_j (\frac{x_{iu}}{\sum_k x_{ku}} - w_{ij}) \tag{3}$$

The weight change from input i to output j was proportional to the normalized input activity at unit i for pattern u, x_{iu}, minus a weight decay term. In addition to the weight decay, the weight to each unit was constrained to sum to one by a divisive normalization.

The small positive activation of non-winning weight vectors does not guarantee that all weight vectors will eventually participate in the clustering. It causes the non-winning weight vectors to move slowly toward the centroid of the data, and some of the weight vectors may end up oscillating about the centroid without winning the competition for one of the inputs. A bias term was therefore added to cause each output unit to be active approximately the same proportion of the time. The learning rule for the bias to output unit j, b_j, was

$$\Delta b_j = \beta \left(\frac{P}{n} - c_j \right) \tag{4}$$

where P is the number of input patterns, n is the number of output units in one pool, and c_j is the count of wins for output j over the previous P time steps. The bias term was updated at the end of each iteration through the data, with learning rate β. If we define a unit's receptive field as the area of input space to which it responds, then the bias term acts to expand the receptive fields of units that tend to be inactive, and shrink the receptive fields of units that are active more often than the others. There is some justification for activity dependent modification of receptive field size of cortical neurons (eg. Jenkins, Merzenich, & Recanzone, 1990; Kaas, 1991). An alternative way to normalize responses is through multiplicative scaling of the synaptic weights (Turrigiano et al., 1998).

One face image was input to the system per time step, so the face patterns, u, can also be indexed by the time step, t. The temporal smoothing was subject to reset based on discontinuities in optic flow, which insured that there was no temporal smoothing across input images with large changes. Optic flow between image pairs was calculated using a simple gradient-based flow estimator (Bergen & Hingorani, 1990). When the summed lengths of the optic flow vectors for sequential image pairs exceeded a threshold of $\gamma = 25$, $\overline{y}$ was initialized to y.[1] 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 components analysis. It has been shown that competitive Hebbian learning can find the first N principal components of the input data, where N is the number of output units (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. Due to the linear transfer function, passing the output activity through a temporal filter is equivalent to passing the input through the temporal filter. Competitive Hebbian learning can thus find the principal components of this spatio-temporal input signal.

2.3 Temporal association in an attractor network

The lateral interconnections in the output layer formed an attractor network. After the feed-forward connections were established in the first layer using competitive learning, 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 (Equation 1) on the

[1] This initialization is not strictly required for the success of such unsupervised learning algorithms because of the low probability of any specific pair of adjacent images of different individuals relative to the probability of adjacent images of the same individual (cf. Wallis & Baddeley 1997). However, we chose not to ignore the transitions between individuals since there are internal cues to these transitions such as eye movements, motion, and longer temporal delays.

unit activities, 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 algorithm:

$$W_{ij} = \frac{1}{N}\sum_{t=1}^{P}(y_i^t - y^0)(y_j^t - y^0) \tag{5}$$

where N is the number of units, P is the number of patterns, and y^0 is mean activity over all of the units. Replacing y_i^t with the activity trace $\overline{y_i}^{(t)}$ defined in Equation 1, we obtain

$$W_{ij} = \frac{1}{N}\sum_{t=1}^{P}\left((1-\lambda)y_i^t + \lambda\overline{y_i}^{(t-1)} - y^0\right)\left((1-\lambda)y_j^t + \lambda\overline{y_j}^{(t-1)} - y^0\right) \tag{6}$$

Substituting $y^0 = \lambda y^0 + (1-\lambda)y^0$ and multiplying out the terms produces the following learning rule:

$$\begin{aligned} W_{ij} = \frac{1}{N}\sum_{t=1}^{P}\Big(& (1-\lambda)^2(y_i^t - y^0)(y_j^t - y^0) \\ & +\lambda(1-\lambda)\left[(y_i^t - y^0)(\overline{y_j}^{(t-1)} - y^0) + (\overline{y_i}^{(t-1)} - y^0)(y_j^t - y^0)\right] \\ & +\lambda^2\left[(\overline{y_i}^{(t-1)} - y^0)(\overline{y_j}^{(t-1)} - y^0)\right]\Big) \end{aligned} \tag{7}$$

This learning rule is a generalization of an attractor network learning rule that has been shown to produce correlated attractors 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):

$$V_i(t+\delta t) = \phi\left[\sum W_{ij}V_j(t) - \theta\right] \tag{8}$$

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 $\lambda = 0.5$.

The learning rule in Griniasty, Tsodyks, and Amit (1993) is presented in Equation 9 for comparison. The Griniasty et. al. learning rule associates first neighbors in the pattern sequence, whereas the learning rule in 7 has a longer memory. The weights in 9 are a function of the *discrete* activities at t and $t-1$, whereas the weights in 7 are a function of the current input and the activity *history* at time $t-1$.

$$W_{ij} = \frac{1}{N}\sum_{t=1}^{P}(y_i^t - y^0)(y_j^t - y^0) + a\left[(y_i^{t+1} - y^0)(y_j^t - y^0) + (y_i^t - y^0)(y_j^{t+1} - y^0)\right] \tag{9}$$

The weight structure and fixed points of an attractor network trained with Equation 7 are illustrated in Figures 4 and 5 using an idealized data set in order to facilitate visualization. The fixed points for the real face data will be illustrated later, in Section 2.4. The idealized data set contained 25 input patterns, where each pattern was coded by activity in a single bit (Figure 4, Top).

The patterns represented 5 individuals with 5 views each (a - e). 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 decrease with the distance between the patterns in the input sequence. The bottom graphs show the sustained patterns of activity in the attractor network for each input pattern. Unlike the standard Hopfield net, in which the objective is to obtain sustained activity patterns that are identical to the input patterns, the objective here is to have a many-to-one mapping from the five views of an individual to a single pattern of sustained activity. Note that the same pattern of activity is obtained no matter which of the 5 views of the individual is input to the network. For this simplified representation, the attractor network produces responses that are entirely viewpoint invariant. The fixed points in this demonstration are the conjunctions of the input activities for each individual view.

Figure 5 shows the weight matrix for different values of the temporal filter, λ.[2] As λ increases, a larger range of views contain positive weights. The figure also gives the fixed points for each input pattern. For $\lambda = 0.25$, 2 to 3 views are associated into the same basin of attraction. For $\lambda = 0.4$, there are positive connections between only a subset of the views for each face, yet this weight matrix is sufficient to associate all five views into the same basin of attraction. A rigorous numerical analysis of the mean field equations and fixed points of a related weight matrix can be found in Parga & Rolls (in press).

2.4 Simulation Results

Sequences of graylevel face images were presented to the network in order as each subject changed pose. Faces rotated from left to right and right to left in alternate sweeps. The feedforward and the lateral connections were trained successively. The feedforward connections were updated by the learning rule in Equations 1-3, with $\lambda = 0.5$. Competitive interactions were among two pools of 35 units so that there were two active outputs for each input pattern. The two competitive pools created two samples of image clustering, which provided additional information on relationships between images. Images could be associated by both clusters, one, or neither, and images that were never clustered together could share a common clustering partner.

After training the feedforward connections, the representation of each face was a sparse representation consisting of the two active output units out of the total of 70 complex pattern units. "Pose tuning" of the feedforward system was assessed by comparing correlations in the network outputs for different views of the same face to correlations across faces of different people. Mean correlations for different views of the same face were obtained for each possible change in pose by calculating mean correlation in feedforward outputs across all four 15^0 changes in pose, three 30^0 changes in pose, and so forth. Mean correlations *across* faces for the same changes in pose were obtained by calculating mean correlation in feedforward outputs for different subjects across all 15^0 changes in pose, 30^0 changes in pose, and so forth.

Figure 6 (Top Left) shows pose tuning 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 discriminability of the feedforward output for same-face versus different-face was measured by calculating the receiver-operator-characteristic (ROC) curve for the distributions of same-face and different-face output correlations. An ROC curve plots the proportion of hits against the proportion of false alarms (FA's) for choosing between two distributions at different choices

[2]The half-life, h, of the temporal filter is related to λ by $\lambda^h = 0.5$ (Stone, 1996). For $\lambda = 0.5$, the activity at time t is reduced by 50% after one time step.

of the acceptance criteria. The area under the ROC measures the discriminability of the two distributions, ranging from 0.5 for fully overlapping distributions to 1.0 for distributions with zero overlap in the tails. Figure 6 (Top Right) shows the ROC curves and areas under the ROC for feedforward output correlations with $\lambda = 0.5$ and $\lambda = 0.0$. The temporal filter increased the discriminability of the feedforward outputs.

Test image results were obtained by alternately training on four poses and testing on the fifth, and then averaging across all test cases. Test images produced a similar pattern of results, which are presented in the bottom of Figure 6.

The feedforward system 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 output units were updated by the learning rule in Equation 7. 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. The sustained patterns of activity comprised the representation of a face in the attractor network component of the model. Following learning, these patterns of sustained activity were approximately viewpoint invariant.

Figure 7 shows pose tuning and ROC curves for the sustained patterns of activity in the attractor network. The graphs compare activity correlations obtained using five values of λ in Equation 7. Note that $\lambda = 0$ corresponds to a standard Hebbian learning rule. The contribution of the feedforward system and the attractor network to the overall viewpoint invariance of the system are compared in Table 1. Temporal associations in the feedforward connections and the lateral connections both contributed to the viewpoint invariance of the sustained activity patterns of the system.

Figure 8 shows the activity in network layer two for 25 of the 100 graylevel face images, consisting of five poses of five individuals. Face representations following training of the feedforward connections only with $\lambda = 0$ (top) are contrasted with face representations obtained when the feedforward connections were trained with $\lambda = 0.5$ (middle), and with the face representations in the attractor network, in which both the feedforward and lateral connections were trained with $\lambda = 0.5$. Competitive Hebbian learning without the temporal lowpass filter frequently included neighboring poses of an individual in a cluster, but the number of views of an individual within the same cluster did not exceed two, and the clusters included images of other individuals as well. The temporal lowpass filter increased the number of views of an individual within a cluster. Note however, that for individuals 4 and 5, the representation of views *a* and *b* are not correlated with that of views *d* and *e*. The attractor network of the bottom plot was trained on the face codes shown in the middle plot, with $\lambda = 0.5$. The attractor network increased the correlation in face codes for different views of an individual. In the sample shown, the representations for individuals 1 - 4 became viewpoint invariant, and the representations for the views of individual 5 became highly correlated. Consistent with the findings of Weinshall & Edelman (1991) for idealized wire-framed objects, units that were active for one view of a face in the input to the attractor network exhibited sustained activity for more views, or all views of that face in the attractor network.

The storage capacity of this attractor network, defined as the maximum number of individual faces that can be stored and retrieved in a view-invariant way, F_{max}, depends on several factors. These include the load parameter, $\frac{P}{N}$, where P is the number of input patterns and N is the number of units, the number of views, s, per individual, and the coding efficiency, or sparseness, y_0. A detailed analysis of the influence of these factors on capacity has been presented elsewhere (Parga & Rolls, in press; See also Gardner, 1988; Tsodyks & Feigel'man, 1988).

We will outline some of these influences here. It has been shown for the autoassociative Hopfield network, for which the number of fixed points equals the number of input patterns, that the

network becomes unstable with $\frac{P}{N} > 0.14$ (Hopfield, 1982). For the present network, we desired one fixed point per individual, where there were $s = 5$ input patterns per individual. Thus the capacity depended on $\frac{F}{N}$, where $F = \frac{P}{s}$ was the number of individuals in the input. The capacity of the attractor network also depended on the sparseness, y_0, since capacity increases as the mean activity level decreases according to $(y_0|ln(y_0)|)^{-1}$ (Gardner, 1988; Tsodyks & Feigel'man, 1988). Specifically, the capacity of attractor networks with $\{0,1\}$ coding and s input patterns per desired memory depends on the number of neurons, N, and the sparseness of the input patterns, y_0, in the following way (Tsodyks & Feigel'man, 1988; Parga & Rolls, in press):

$$\frac{F}{N} \leq \frac{0.2}{s^2 y_0 \ ln\left(\frac{1}{sy_0}\right)} \tag{10}$$

For the network with $N = 70$ units, sparseness $y_0 = 0.029$, and $s = 5$ views per individual, the maximum load ratio was $\frac{F}{N} = 0.14$, and the maximum number of individuals that can be stored in separate basin of attraction was $F_{max} = 10$.

Since storage capacity in the attractor network depends on coding efficiency, the proportion of active input units per pattern, the attractor network component of the model required its input representations to be sparse. Sparse inputs may be an appropriate assumption, given the sparseness of responses reported in V4 (Gallant, 1994) and area TE, a posterior IT region which projects to the anterior IT regions where transformation invariant responses can be found (Tanaka, 1993). The representations of faces in the attractor network itself were less sparse than its input, with a mean unit activity of 0.19 for each face, compared to 0.03 for its input, and each unit participated in the coding of 13 of the 100 faces on average in the attractor network, compared to 3 faces for its input. The coding levels in the attractor network were consistent with the sparse-distributed face coding reported in IT (Abbott, Rolls, & Tovee, 1996; Young & Yemane, 1992).

We evaluated face recognition performance of the attractor network using a nearest neighbor classifier on the sustained activity patterns at several loading levels. Table 2 gives percent correct recognition performance of the sustained activity patterns in the network trained on real face data. Test patterns were assigned the class of the pattern that was closest in Euclidean distance. Each pattern was taken in turn as a test pattern and compared to the other 99, and then a mean was taken across the 100 test cases. Classification performance depended on the load parameter, $\frac{F}{N}$. Performance was quite good when $\frac{F}{N} << 0.14$, and decreased as $\frac{F}{N}$ increased beyond this value. Classification errors occurred when two or more individuals shared a single basin of attraction.

Classification performance of the network for $F = 10$ was below 100% because not all fixed points were found. The set of input patterns did cover all 10 basins of attraction. Since the input patterns (the outputs of the feedforward system) were driven by real face images, the input patterns were not constrained to be orthogonal. When the input patterns were orthogonal, such as the idealized data in Figure 4 in which each input was coded by activity in a different unit, then all fixed points were found for $F = F_{max}$ individuals, and classification performance was 100%.

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 of faces could be developed from visual experience through unsupervised learning.

The inputs to the model were similar to the responses of V1 complex cells, and the goal was to apply unsupervised learning mechanisms to 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. Human subjects were better able to recognize famous faces when the faces were presented in video sequences, as compared to an array of static views (Lander & Bruce, 1997). Recognition of novel views of unfamiliar faces was superior when the faces were presented in continuous motion during learning (Pike, Kemp, Towell, & Phillips, 1997). Stone (1998) found that recognition rates for rotating amoeboid objects decreased, and reaction times increased when the temporal order of the image sequence was reversed in testing relative to the order during learning. The dynamic signal therefore contributed to the object representation beyond providing structure-from-motion. This model in this paper presented a means by which temporal information can be incorporated in the representation of a face.

Related models that have been developed independently support the results presented in this paper. Wallis and Rolls (Wallis & Rolls, 1997) trained a hierarchical feedforward system using Hebbian learning and the temporal activity trace of Equation 1. Their system successfully learned translation invariant representations of seven faces, and rotation invariant representations of three faces. Parga and Rolls (in press) presented a detailed analysis of the phase transitions and capacity of an attractor network related to the recurrent layer of the present network. Their work focused on the thermodynamic properties of this attractor network, using a predefined coupling matrix and idealized stimuli. Our work extends this analysis to the learning mechanisms that could give rise to such a weight matrix, and implements them in a system taking real images of faces as input.

The feedforward processing in this model was related to spatio-temporal principal components analysis of the Gabor filter representation. It has been shown that competitive Hebbian learning finds the principal components of the input data (Oja, 1989; Sanger, 1989). 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.

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 not require explicit internal 3-dimensional models, as was previously assumed, and recognition of novel views may instead be accomplished by linear (Ullman & Basri, 1991) or nonlinear combination of stored 2-D views (Poggio & Edelman, 1990; Bulthoff, Edelman & Tarr, 1995). Such view-based representations may be particularly relevant for face processing, given the recent psychophysical evidence for face representations based on low-level filter outputs (Biederman, in press; Bruce, in press).

Further support for view-based representations comes from a related model that simulated "mental rotation" response curves in a system that stored multiple 2-dimensional views and their temporal associations (Weinshall & Edelman, 1991). Weinshall and Edelman trained a 2 layer network to store individual views of wire-framed objects, and then updated lateral connections in the output layer with Hebbian learning as the input object rotated through different views. The

strength of the association was proportional to the estimated strength of the perceived apparent motion if the 2 views were presented in succession to a human subject. After training the lateral connections, one view of an object was presented and the output activity was iterated until all of the units for that object were active. When views were presented that differed from the training views, correlation in output ensemble activity decreased linearly as a function of rotation angle from the trained view, mimicking the linear increase in human response times that has been taken has evidence for mental rotation of an internal 3-D model (Shepard & Cooper 1982).

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. Our 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. We thank Tomaso Poggio, James Stone, and Laurenz Wiskott for valuable discussions on earlier drafts of this paper.

References

Abbott, L., Rolls E., & Tovee, M. (1996). Representational capacity of face coding in monkeys. *Cerebral Cortex* 6(3):498-505.

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*,817-823, Cambridge, MA: MIT Press.

Becker, S. (1993). Learning to categorize objects using temporal coherence. In S. Hanson, J. Cowan, & C. Giles, (Eds.)*Advances in Neural Information Processing Systems 5*, 361-368, San Mateo, CA: Morgan Kaufman.

Becker, S. (in press). Implicit learning in 3-D object recognition: The importance of temporal coherence. *Neural Computation.*

Beymer, D. (1994). Face recognition under varying pose. In *Proceedings of the 1994 IEEE Computer Society Conference 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, P. Phillips, V. Bruce, F. Fogelman-Soulie, & T. Huang, Eds., *Face Recognition: From Theory to Applications.* Springer-Verlag.

Bruce, V. (1988). *Recognising Faces.* London: Lawrence Erlbaum Assoc.

Bruce, V. (in press). Human face perception and identification. In H. Wechsler, P. Phillips, V. Bruce, F. Fogelman-Soulie, & T. Huang, 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.

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.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Comp.* 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.

Gallant, J.L., Connor, C.E., & Van Essen, D.C. (1994). Responses of visual cortical neurons in a monkey freely viewing natural scenes *Society for Neuroscience Abstracts* 838.

Gardner, E. (1988). The space of interactions in neural network models. *Journal of Physics A: Math. Gen.* 21: 257-270.

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

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: Part 1. Parallel

development and coding of neural feature detectors. *Biological Cybernetics* 23: 121-134.

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.

Heeger, D. (1991). Nonlinear model of neural responses in cat visual cortex. In M. Landy & J. Movshon, (Eds.) *Computational Models of Visual Processing*, 119-133, 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.

Horn, B. & Schunk, B. (1981). Determining optical flow. *Artificial Intelligence* 17: 185 - 203.

Jenkins, W.M., Merzenich, M.M., & Recanzone, G. (1990). Neocortical representational dynamics in adult primates: implications for neuropsychology. *Neuropsychologia* 28(6): 573-84.

Kaas, J.H. (1991). Plasticity of sensory and motor maps in adult mammals. *Annual Review of Neuroscience* 14: 137-67.

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

Lander, K. & Bruce, V. (1997). The role of movement in the recognition of famous faces. Poster presentation, *NATO ASI on Face Recognition: From Theory to Applications. Stirling, Scotland.* (Submitted for journal publication.)

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

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

Montague, R., Gally, J., & Edelman, G., (1991). Spatial signaling in the development and function of neural connections. Cerebral Cortex: 1: 199-220.

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

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

Parga N. & Rolls, E. (in press). Transform invariant recognition by association in a recurrent network. *Neural Computation.*

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

Pike, G.E., Kemp, R.I., Towell, N.A.s and Phillips, K.C. (1997). Recognizing moving faces: The relative contribution of motion and perspective view information. *Visual Cognition* 4(4): 409-437.

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

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. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 335(1273): 11-20.

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.

Shepard, R.N. & Cooper, L.A. (1982). *Mental Images and their Transformations.* Cambridge, MA: MIT Press.

Stone, J.V. (1996). Learning perceptually salient visual parameters using spatiotemporal smoothness constraints. *Neural Computation* 8(7): 1463-92.

Stone, J.V. (1998). Object Recognition Using Spatiotemporal Signatures. *Vision Research* 38(7): 947-951.

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

Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science* 262: 685-688.

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

Turk, M., & Pentland, A. (1991). Eigenfaces for Recognition. *J. Cog. Neurosci.* 3(1):71-86.

Turrigiano GG; Leslie KR; Desai NS; Rutherford LC; Nelson SB. (1998). Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature* 391(6670): 892-6.

Ullman, S. & Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13(10): 992-1006.

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

Wallis, G. & Baddeley, P. (1997). Optimal, unsupervised learning in invariant object recognition. *Neural Computation* 9(4): 883-94.

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

Young, M. & Yemane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science* 256: 1327-1331.

	λ_1	
λ_2	0	0.5
0	.70	.90
0.5	.84	.98

Table 1: Contribution of the feedforward connections and the attractor network to viewpoint invariance of the complete system. Area under the ROC for the sustained activity patterns in network layer 2 is given with and without the temporal activity trace in during learning in the feedforward connections (λ_1) and in the attractor network (λ_2).

		Attractor Network			Eigenfaces
F	P	N	$\frac{F}{N}$	%Correct	% Correct
5	25	70	.07	100	100
10	50	70	.14	90	90
20	100	70	.29	61	87

Table 2: Nearest neighbor classification performance of the attractor network. F: Number of individuals; P: Number of input patterns; N: Number of units. Classification performance is presented for three values of the load parameter, $\frac{F}{N}$. Results are compared to Eigenfaces for the same subset of faces. Classification performance of the attractor network is good when $\frac{F}{N} < 0.14$

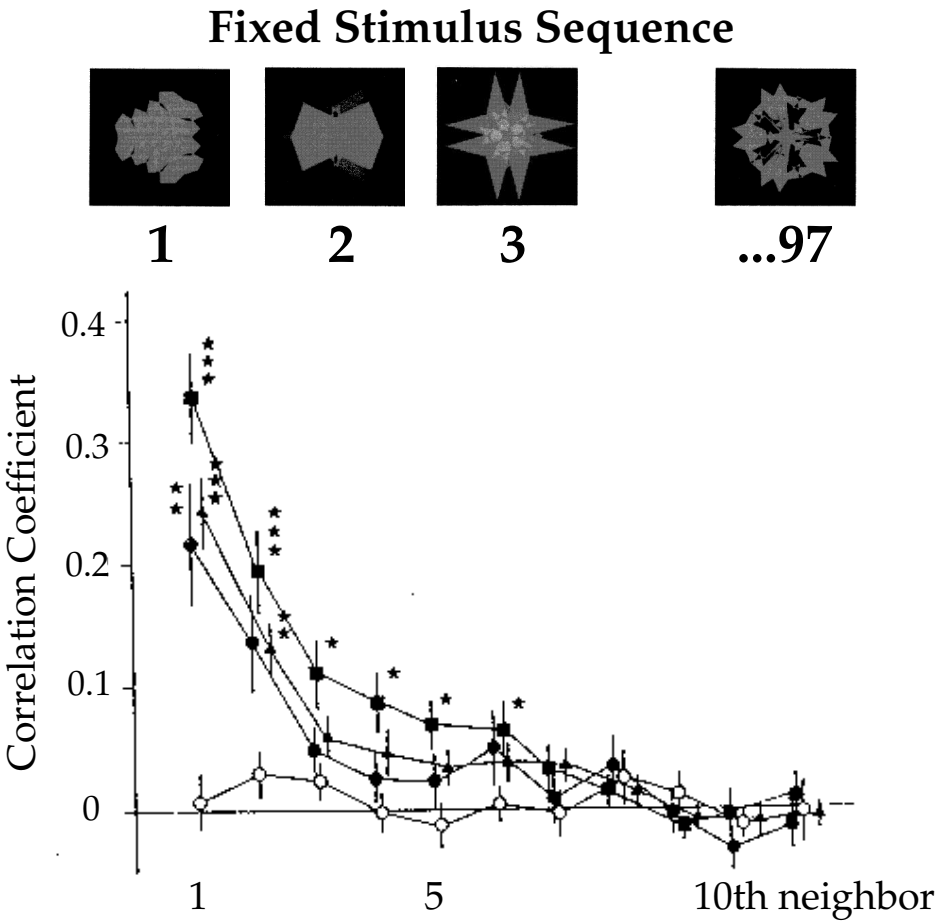

Figure 1: Evidence of temporal associations in IT. Top: Samples of the 97 fractal pattern stimuli in the fixed training sequence. Bottom: Autocorrelograms on the sustained firing rates of AIT cells along the serial position number of the stimuli. Abscissa is the relative position of the patterns in the training sequence, where patterns n,n+1 are first neighbors, and patterns n,n+2 are second neighbors. 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.

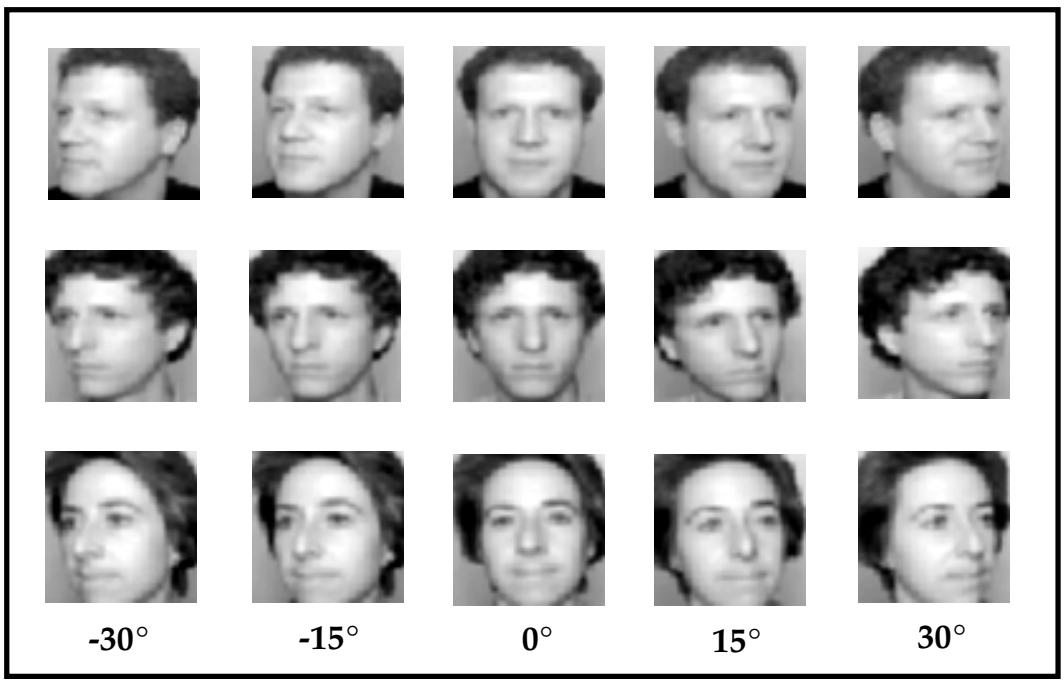

Figure 2: Sample of the 100 images used in the simulation. Image set provided by David Beymer (Beymer, 1994).

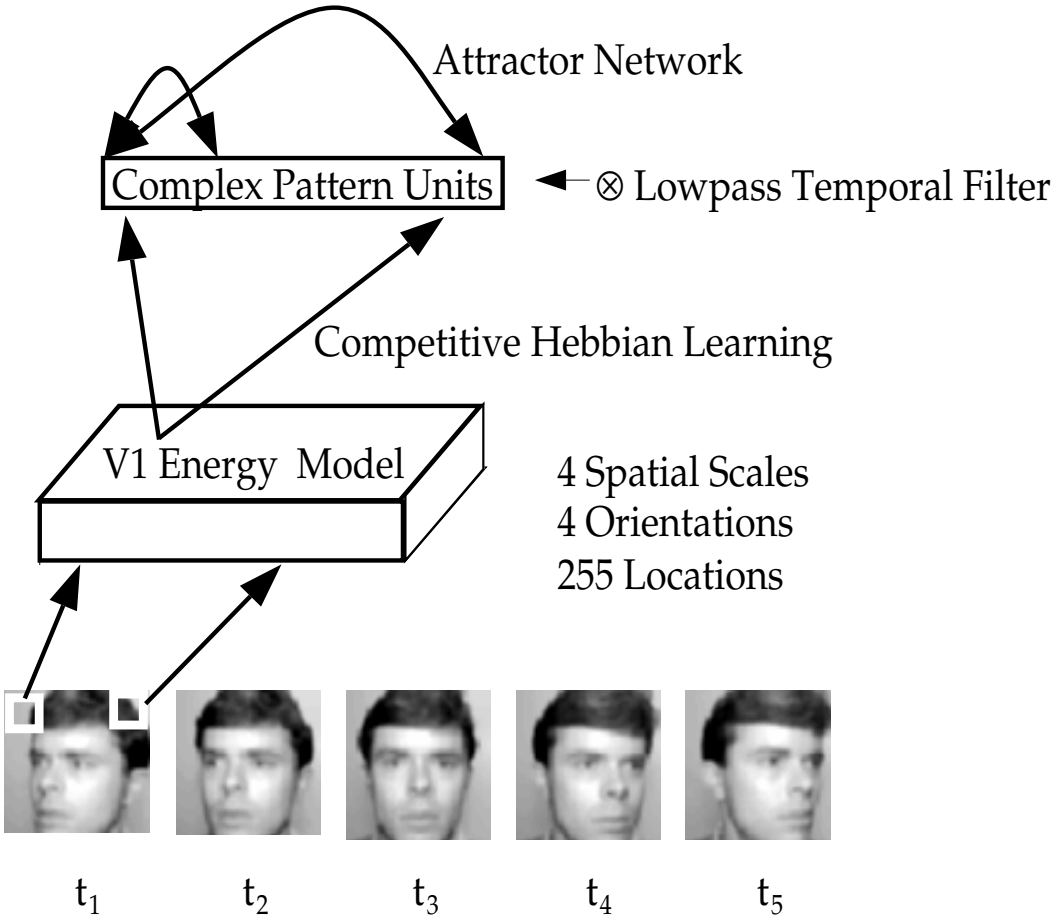

Figure 3: Model architecture.

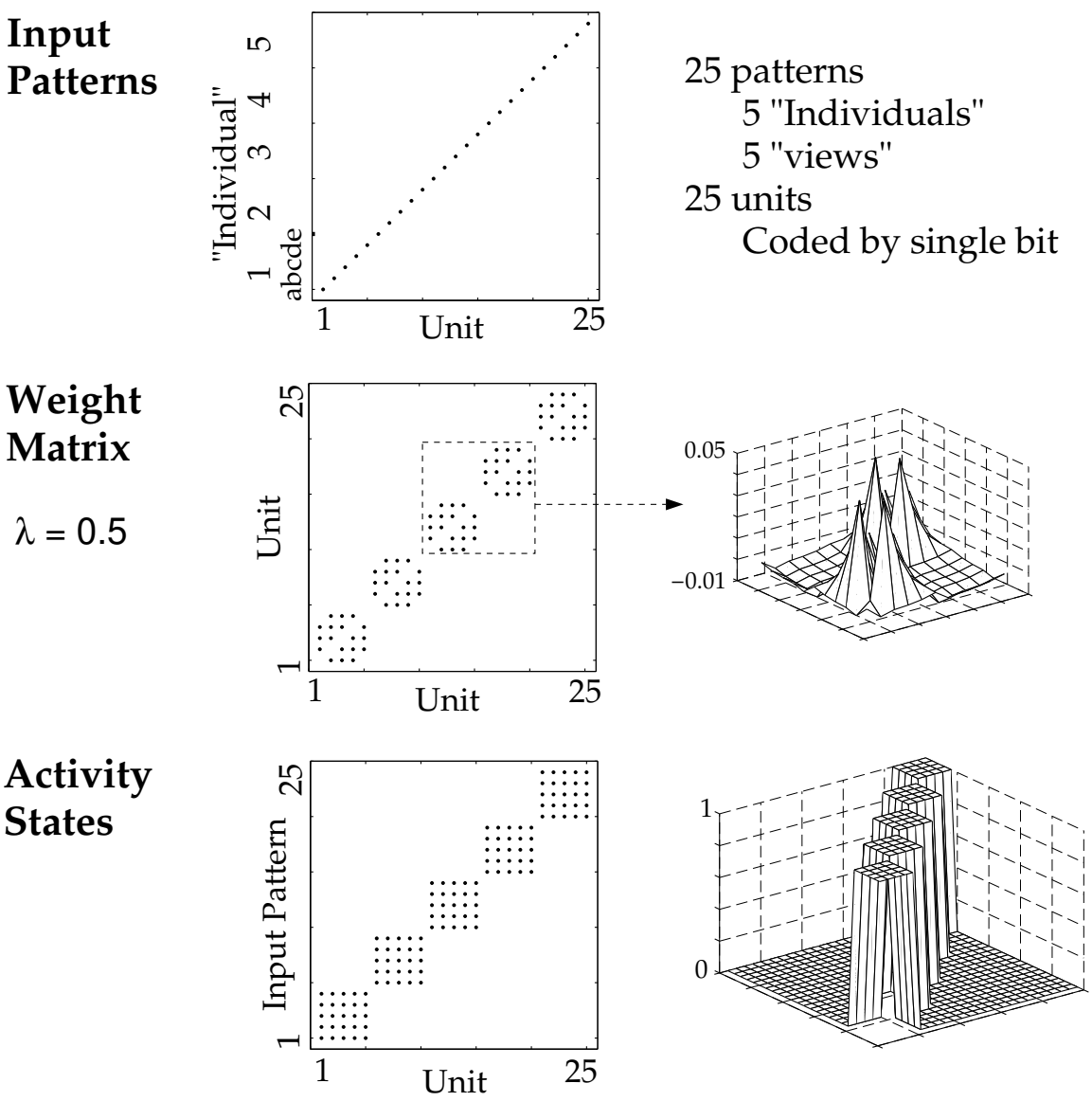

Figure 4: Demonstration of attractor network with idealized data. Top: Idealized data set. The patterns consist of 5 "individuals" (1,2,3,4,5) with five "views" each (a,b,c,d,e), and are each coded by activity in 1 of the 25 units. Center: The weight matrix obtained with equation 3. Dots show the locations 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.

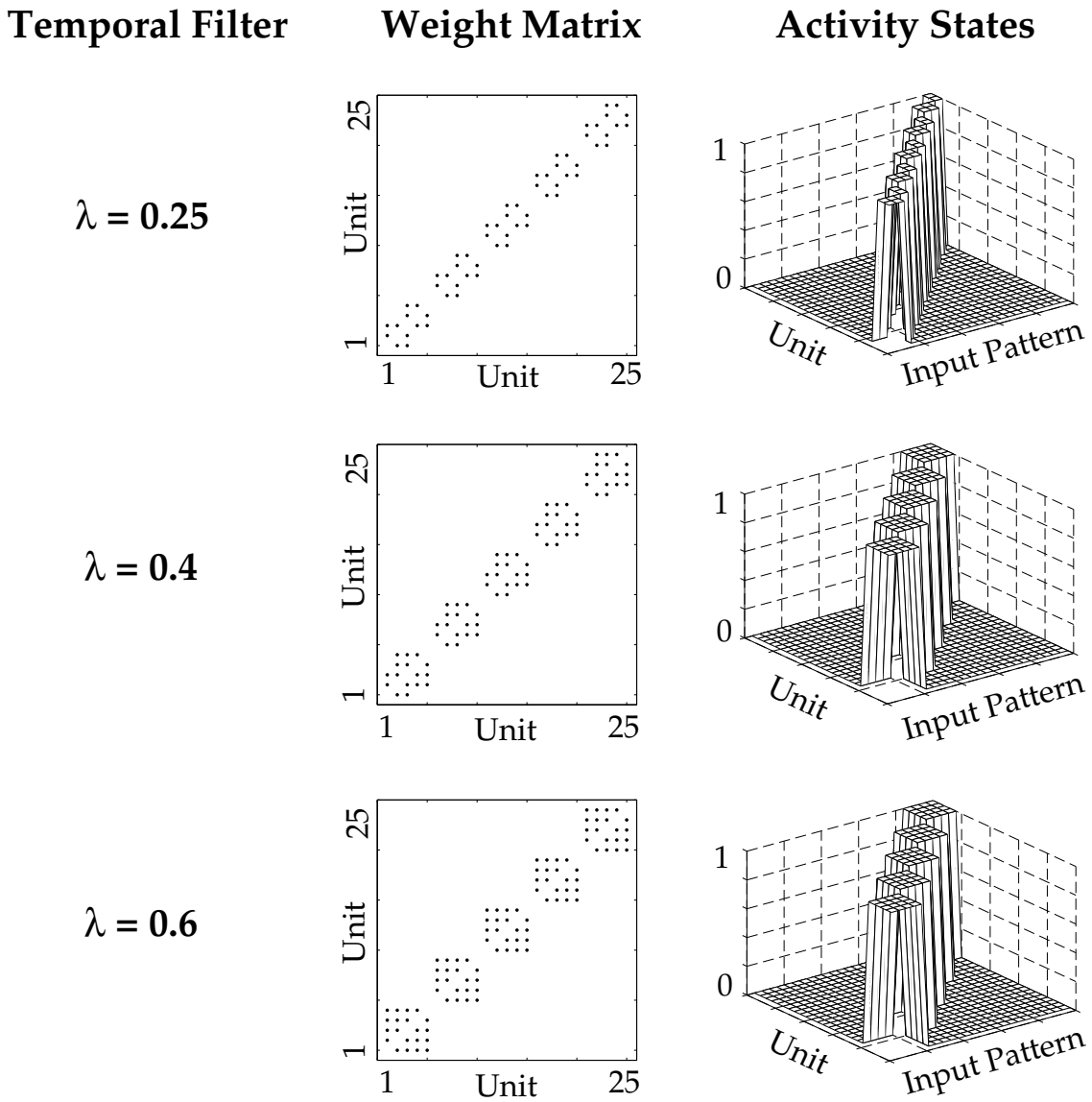

Figure 5: Weight matrix (left) and fixed points (right) for three values of the temporal filter, λ. Dots show locations of positive weights. Unit activities are plotted for each of the 25 input patterns of the simplified data.

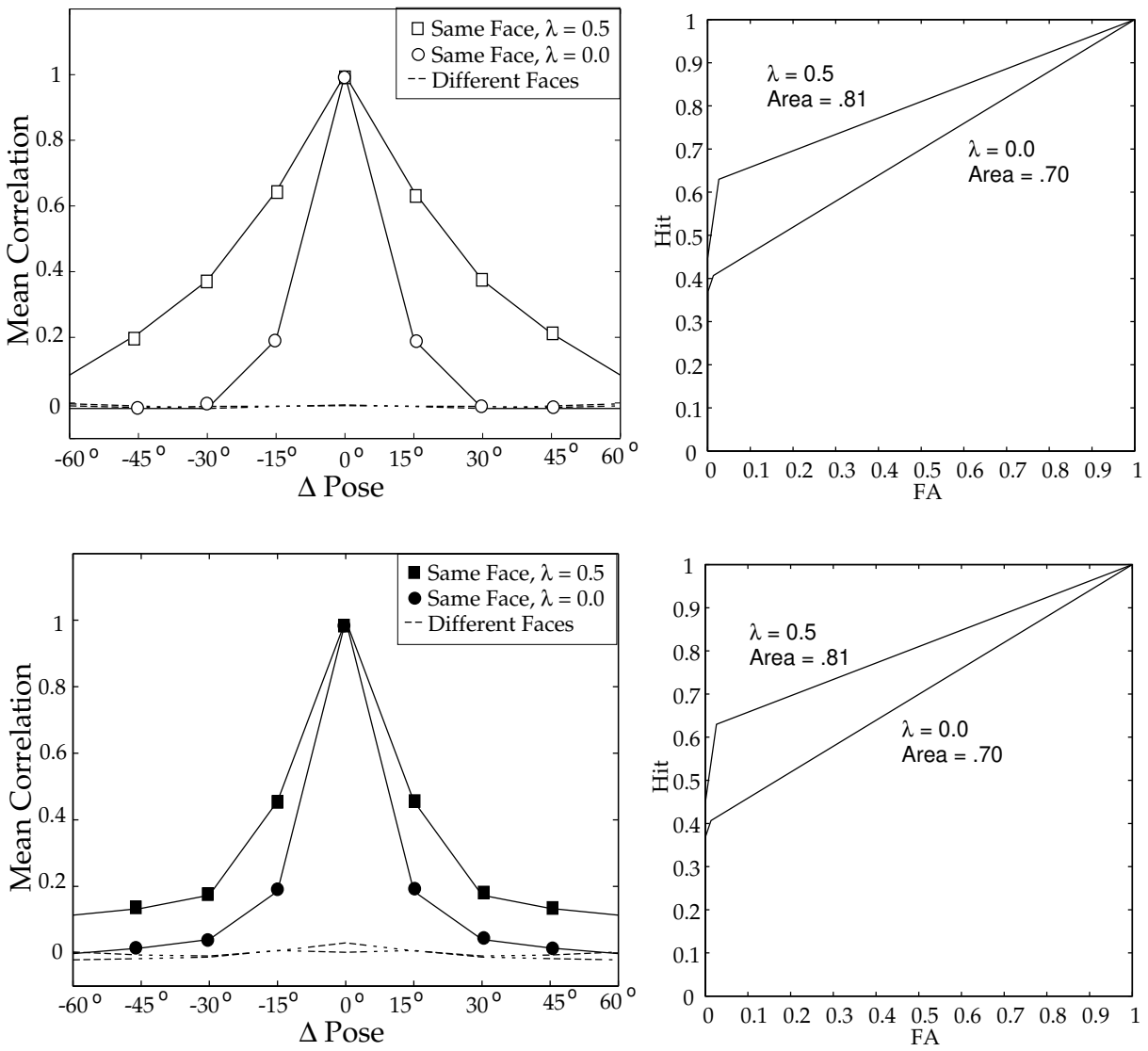

Figure 6: Pose tuning and ROC curves of the feedforward system for training images (top) and test images (bottom). Left: Mean correlations of the feedforward system outputs for pairs of face images are presented by change in pose. Correlations across different views of the same face (–) are compared to correlations across different faces (––) for two values of the temporal trace parameter $\lambda = 0.5$ and $\lambda = 0$. Right: ROC curves and area under the ROC for same face vs. different face discrimination of the feedforward system outputs for training images (top) and test images (bottom).

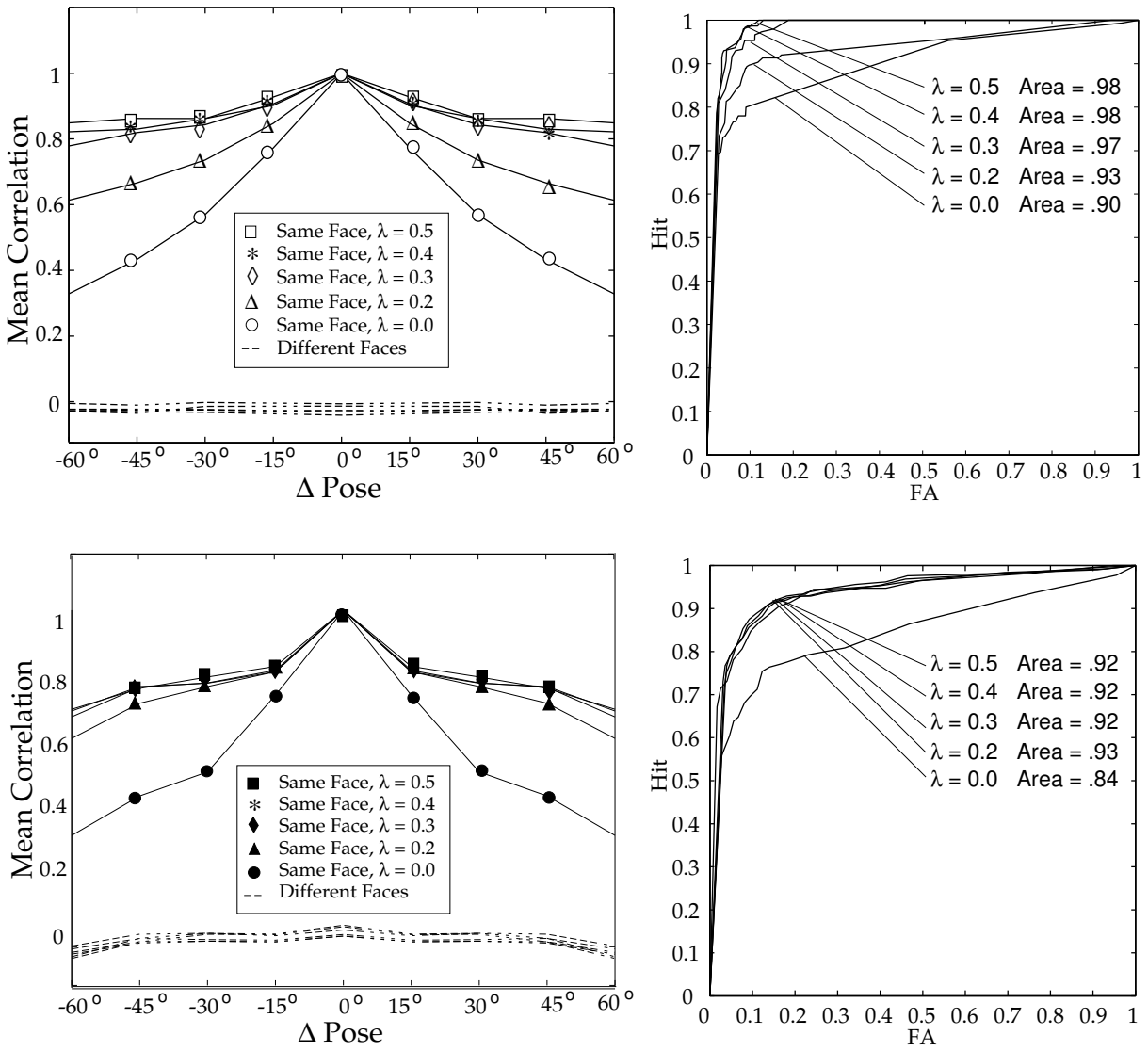

Figure 7: Pose tuning and ROC curves of the attractor network for training images (top) and test images (bottom). Left: Mean correlations in sustained activity patterns in the attractor network for pairs of face images are presented by change in pose. Correlations across different views of the same face (–) are compared to correlations across different faces (– –) for five values of the temporal trace parameter λ. Right: ROC curves and area under the ROC for same face vs different face discrimination of the sustained activity patterns for training images (top) and test images (bottom).

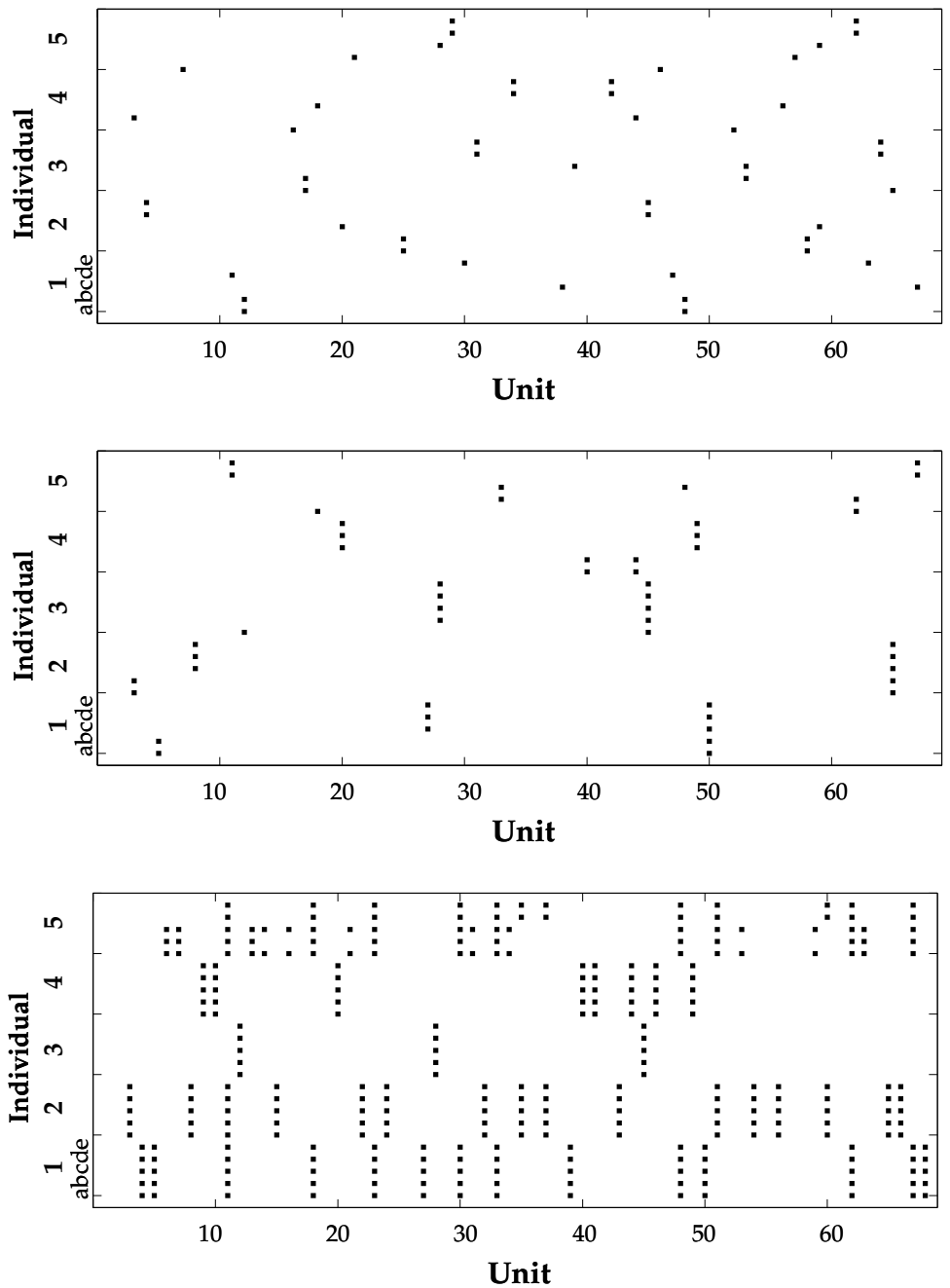

Figure 8: Coding of real face image data. Top: Coding of 5 faces in network layer 2 following training of the feedforward connections only, with no temporal lowpass filter ($\lambda = 0.$) The vertical axis is the input image, with the five poses of each individual labeled a,b,c,d,e. The two active units for each input image are indicated on the horizontal axis. Middle: Coding of the same five faces following training of the feedforward connections with $\lambda = 0.5$. Bottom: Sustained patterns of activity in the attractor network for the same five faces, where both the feedforward and the lateral connections were trained with $\lambda = 0.5$.