In press: *The Neurobiology of Computation: Proceedings of the Annual Computational Neuroscience Meeting.* J.M. Bower, ed. Kluwer Academic Publishers, Boston.

UNSUPERVISED LEARNING OF INVARIANT REPRESENTATIONS OF FACES THROUGH TEMPORAL ASSOCIATION

Marian Stewart Bartlett[*,‡]
Terrence J. Sejnowski[†,‡]

marni@salk.edu, terry@salk.edu

[*] *Departments of Cognitive Science and Psychology, UCSD*
[†] *Howard Hughes Medical Institute*
[‡] *The Salk Institute, La Jolla, CA, 92037*

Abstract

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. Recognizing an object or a face despite these image changes is a challenging problem for computer vision systems, yet we perform the task quickly and easily. This simulation investigates the ability of an unsupervised learning mechanism to acquire representations that are tolerant to such changes in the image. The learning mechanism finds these representations by capturing temporal relationships between 2-D patterns. Previous models of temporal association learning have used idealized input representations. The input to this model consists of graylevel images of faces. A two-layer network learned face representations that incorporated changes of pose up to $\pm 30°$. A second network learned representations that were independent of facial expression.

Introduction

One of the greatest challenges in visual recognition of objects or faces is that the projected image can vary substantially with changes in viewing conditions. In normal visual experience, however, these different views tend to appear in close temporal proximity. Unsupervised learning can find invariant representations by capitalizing on this dynamic information. Capturing the temporal relationships among patterns is a way to automatically associate different views of an object without requiring complex geometrical transformations or three dimensional structural descriptions [1].

Temporal association may be a fundamental component of visual processing in the temporal lobe. Cells in the anterior inferior temporal lobe will adjust their receptive fields so that they respond to temporally contiguous inputs [2]. A temporal window for Hebbian learning could be provided by the long open-time of the NMDA channel [3], a hysteresis in neural activity caused by reciprocal connections between cortical regions [4], or the release of a chemical signal following activity such as nitric oxide [5].

This simulation investigates the capability of such Hebbian learning mechanisms to acquire transformation invariant representations of complex objects such as faces. These mechanisms have been previously tested with idealized input representations with little or no crosstalk on the connections [6, 7, 4]. In order to understand the capabilities of temporal association learning, it is important to evaluate it using complex, realistic stimuli.

Network Architecture

We tested the temporal association learning mechanism on a very simple architecture (Figure 1). We used a feed-forward network with two layers of units. There were 400 input units and ten output (representation) units. The input layer was fully connected to the output layer and there was winner-take-all competition in the output layer. We used a linear transfer function, and the total weight coming into each output unit was constrained to sum to one. At each time step t, the network took one 20 x 20 graylevel image as input.

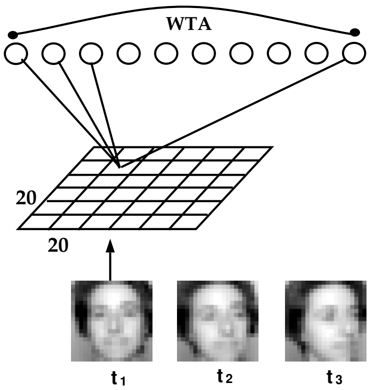

Figure 1: Network architecture. Images at resolution used in the simulations.

Temporal Association Learning

The weight update rule is based on the Competitive Learning Rule [8, 9]. Let α be the learning rate, x_{ik} be the value of input unit i for pattern k, and t_k be the total amount of input activation for pattern k. The weight update rule is

$$\Delta w_{ij} = \begin{cases} \alpha \frac{x_{ik}}{t_k} - \alpha w_{ij} & \text{if } winner = j \\ 0.1(\alpha \frac{x_{ik}}{t_k} - \alpha w_{ij}) & \text{if } winner \neq j \end{cases}$$

We introduce a temporal manipulation into the competition phase. Let y_j be the activation of output j, computed by a weighted sum of the inputs. After Foldiak [6], the winning unit i at time t is determined by the trace of the activation:[1]

$$winner = max_j[y_j^t]$$
$$y_j^t = (1-\lambda)y_j^{t-1} + \lambda y_j$$

The Competitive Learning Rule alone, without the temporal manipulation, will partition the set of inputs into roughly equal groups by spatial similarity. The resulting weights to each output unit are proportional to the probability that a given input unit is active when that unit wins [8]. The temporal manipulation allows temporal association to influence these partitions. The winning unit in the current time step has a competitive advantage for recruiting the pattern in the next time step. This learning rule therefore partitions the input by a combination of spatial similarity and temporal proximity, where λ determines the relative influence of the two factors.

Simulation 1: Learning Invariance to Changes in Pose

We first tested the ability of this learning algorithm to develop representations of faces that were independent of pose. The inputs were graylevel face images provided by David Beymer at the MIT Media Lab [10]. We used images of ten individuals at each of five different angles of view ($0°$,± 15 °, and $\pm 30°$), for a total of fifty stimuli (Figure 2). A single window based on the eye and mouth positions in the frontal view was used for cropping and scaling the other images in each sequence. The faces were reduced to 20 x 20 pixels, producing a 400 dimensional input vector, and each image was normalized for luminance.

The learning was performed in two stages. In the first stage, the network was exposed only to the ten frontal view faces in order to associate each face with a different output unit. Once this initial correspondence was established, the training set was slowly expanded to include variations in pose. Images were presented in sequence beginning with the frontal view. The trace function was reset between sequences.

Results

The network stabilized after 100 training epochs. Network classification was assessed by presenting each image individually, without the activation trace, and recording the output unit with the highest activation. The network response was considered "correct" if the winning output unit is the one corresponding to the frontal view of that subject. Figure 3a compares invariance to pose

[1] Representation units that failed to win for two iterations were given a competitive advantage by increasing slightly the value of y_j^t. This adjustment is equivalent to enlarging the receptive field of that unit [7].

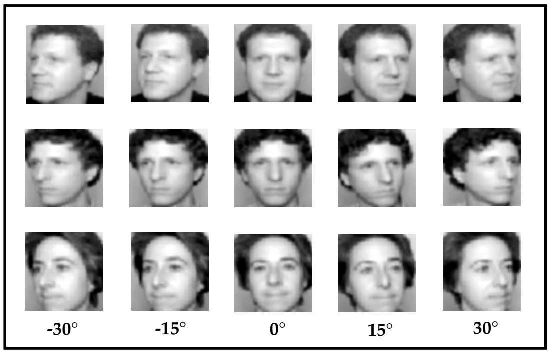

Figure 2: Sample pose sequences. The example set contained ten subjects.

after temporal association learning (dashed line) to baseline performance (solid line), in which the network was trained on the frontal views only. Mean correct classification at each pose is shown, collapsed over the ten subjects in the data set. The graph is analogous to a mean tuning curve for pose.

Temporal association (TA) learning improved the mean classification accuracy of the $\pm 15°$ views from 65% to 90% and increased invariance to the $\pm 30°$ views from 55% to 90%. Performance for the $\pm 15°$ views initially reached 100%, but fell to 95% when the $\pm 30°$ views were added, indicating the beginnings of interference between the patterns.

To test for interpolation between and extrapolation beyond the set of training views, we retrained the network, this time reserving some of the poses as test images. Figure 3b shows an increase in accuracy for the $\pm 30°$ views following training on the $0°$ and $\pm 15°$ views only, and an increase in accuracy for the $\pm 15°$ views when the network was trained on the $0°$ and $\pm 30°$ views only.

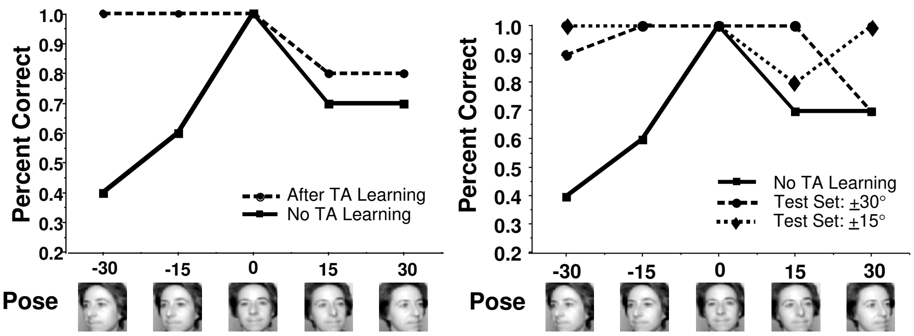

Figure 3: a. Mean tuning curves for pose at baseline and after temporal association (TA) learning. b. Interpolation and extrapolation.

Simulaton 2: Learning Invariance to Facial Expression

The learning rule was also tested for learning representations invariant of facial expression. Changes in facial expression introduce a particular challenge to recognition systems, as they produce a non rigid deformation in the image. The input to the network consisted of ten faces in six sequential stages of an expression, from low to full muscle contraction intensity (Figure 4). If the images included hair, Competitive Learning alone correctly classified all of the images. The task was therefore made more difficult by cropping out the hairline. Training was performed in two phases as above.

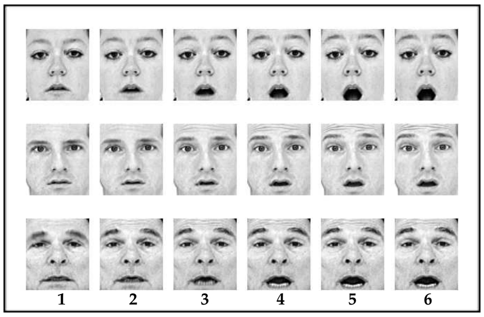

Figure 4: Sample facial expression sequences. Images provided by Paul Ekman and Joe Hager at the Human Interaction Laboratory, UCSF.

Results

Figure 5a shows the increase in invariance to facial expression due to temporal association learning. These results are following exposure to sequences of length 3. The addition of frame 4 to the training sequence caused the network to destabilize, revealing the limits of the range of invariance that this learning method can achieve on this kind of dataset. This network also showed interpolation and extrapolation between trained expression intensities following training on frames 1 and 3 alone (Figure 5b).

Summary and Conclusions

By associating patterns by temporal proximity, our system developed representations of faces with a degree of invariance to changes in pose or changes in facial expression. This simulation demonstrates that unsupervised learning can solve a challenging problem in object recognition, and provides another example of how problems in image understanding can be simplified by taking advantage of dynamic information. This is an idea that has been espoused by the "active vision" approach to computer vision.

The extent of invariance and the number of subjects that this system can tolerate is limited by the redundancy in the input representation. If there is no

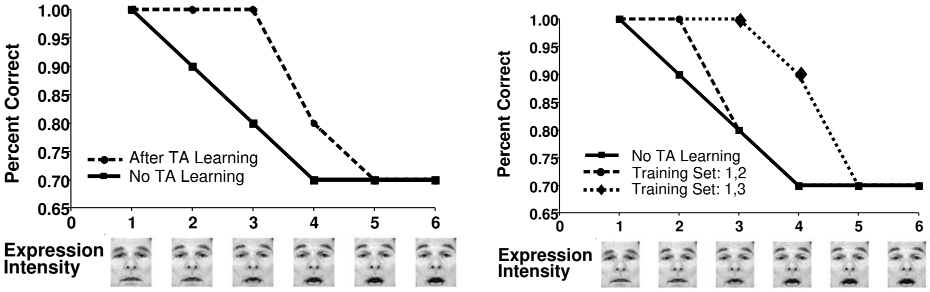

Figure 5: a. Invariance to changes in facial expression before and after temporal association learning on frames 1-3. b. Interpolation and Extrapolation.

redundancy in the input, then there is no limit to the amount of invariance that this system can learn. This points to the importance of intermediate representations with reduced input redundancy, such as principal components or sparse distributed representations [11]. Larger invariances can also be obtained in a hierarchical system that learns new invariances at each level of the hierarchy.

Acknowledgments

This research was supported by Lawrence Livermore National Laboratory, Intra-University Agreement B291436, and Howard Hughes Medical Institute.

References

1. Stryker, M. 1991. Temporal Associations. Nature: 354(14):108-109.

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

3. Rhodes, P. 1992. The long open time of the NMDA channel facilitates the self-organization of invariant object responses in cortex. Society for Neuroscience Abstracts 18:740.

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

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

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

7. Weinshall, D., Edelman, S., & Bulthoff, H. A self-organizing multiple view representation of 3D objects in *Advances in Neural Information Processing Systems No. 2,* D. Touretzky, ed. Cambridge MA: MIT Press, 1990: 274-281.

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

9. Grossberg, S. 1976. Adaptive pattern classification and universal recoding: Part 1. Parallel development and coding of neural feature detectors. Biological Cybernetics: 23:121-134.

10. Beymer, D. 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, 1994: 756-61.

11. Field, D. 1994. What is the goal of sensory coding? Neural Computation: 6(4):559-601.