Communicated by Peter Dayan

Optimal, Unsupervised Learning in Invariant Object Recognition

Guy Wallis
Max-Planck-Institut für biologische Kybernetik, 72076 Tübingen, Germany

Roland Baddeley
Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, UK

A means for establishing transformation-invariant representations of objects is proposed and analyzed, in which different views are associated on the basis of the temporal order of the presentation of these views, as well as their spatial similarity. Assuming knowledge of the distribution of presentation times, an optimal linear learning rule is derived. Simulations of a competitive network trained on a character recognition task are then used to highlight the success of this learning rule in relation to simple Hebbian learning and to show that the theory can give accurate quantitative predictions for the optimal parameters for such networks.

1 Introduction

How might we learn to recognize an object irrespective of the precise relationship between viewer and object, characterized by viewing angle, distance, and translation? If we believe the view-centered approach to object recognition (Tarr & Pinker, 1989; Bülthoff & Edelman, 1992), then the problem becomes one of associating together a series of different views of the object that may share few, if any, of the features supporting the recognition. A broadly tuned feature-based system would be sufficient to perform recognition over small transformations (Poggio & Edelman, 1990), and the necessary receptive fields might be learned via a simple competitive network using Hebbian learning. However, large-shape transformations would require either separate prenormalization for size and translation or separate feature detectors feeding into a final arbitration layer. This then raises the question of how such a final layer might be trained, dismissing the use of any form of supervised training signal.

The use of prenormalization and a final arbiter is in fact in contrast to the evidence we have from the responses of real neurons implicated in object recognition. Invariance seems to be established over a series of processing stages, starting from neurons with restricted receptive fields and culminating in the types of cell responses found in inferior temporal (IT) cortex (Perrett & Oram, 1993; Rolls, 1992). Cells in this region exhibit invariance

to combinations of the types of transformations discussed here (Rolls, 1992; Desimone, 1991; Tanaka, Saito, Fukada, & Moriya, 1991).

Miyashita (1988) has illustrated that arbitrary stimuli can be associated together by a single IT neuron in primates. The key to the training he used was the association of these images in time, not in space. The hypothesis expressed here is that temporal relations in the appearance of object views affect learning. In the real world, we see objects for protracted if variable periods while they undergo any manner of natural transformations, such as when we approach an object or rotate it in our hand. The consistent stream of images serves as a cue that all of these images belong to the same object and that it would be expedient to associate them together.

This idea has been described elsewhere (Edelman & Weinshall, 1991; Földiák, 1991; Wallis & Rolls, 1997); in this article, we aim to build a theoretical framework for analyzing the response of a neuron over time. From this framework we then intend to derive the form of an optimal, local training rule given the general probability function governing the length of each exposure to images of the same object.

2 Mapping the Output of a Neuron to the Optimal Training Signal

The most obvious signal present in a neuron for performing learning is the activity of the neuron itself. Unfortunately, this Hebbian training signal will not typically be optimal for learning-invariant object recognition due to erroneous classifications made by the neuron to spatially similar images from different objects and spatially dissimilar images derived from the same object.

A potentially useful piece of a priori knowledge to overcome this problem is that objects tend to be viewed over variable but extended periods of time. In this work the consequent temporal structure will be used to provide information for improving the use of local neural activity as a training signal. This is achieved here by using a weighted sum of previous neuronal activity rather than by attending to the current neuronal output alone.

Assuming the neuronal output to be a signal described by the function $y(t)$, and the ideal training signal to be $s(t)$, discrepancies between the two signals can be regarded as due to some noise signal $n(t)$. Couched in these terms, the optimal recovery of $s(t)$ from $y(t)$ can be regarded as a classical filtering problem.

The form of the optimal linear filter for retrieving $s(t)$ can be derived by Wiener filtering (Wiener, 1949; Press, 1992).[1] The Wiener filter, $\phi(t)$, gives an optimal estimate, $\tilde{s}(t)$ (in the least-squares sense), of the true signal, $s(t)$, when applied to the noise-contaminated output, $y(t)$. Adopting the con-

[1] There are presumably innumerable, more optimal nonlinear filters, which are beyond the scope of the Wiener filtering theory given here.

vention that capital letters denote functions transformed into the frequency domain, the theorem states:

$$\tilde{S}(f) = Y(f)\Phi(f) \tag{2.1}$$

$$\Phi(f) = \frac{|S(f)|^2}{|S(f)|^2 + |N(f)|^2}\,. \tag{2.2}$$

Hence, the optimal filter $\phi(t)$ can be determined directly from the estimated system noise $n(t)$ and the true signal $s(t)$. The following two sections consider the predicted optimal filter for three different forms of $s(t)$, each describing a different regime for stimulus presentation.

2.1 Fixed-Length Presentation Times. As a simple first case, a set of stimuli are assumed to appear in identical, fixed-length sequences. Possible forms of the ideal training signal and neural output associated with this training regime appear in Figure 1.

The form of $s(t)$ is easily characterized, but there remains the question of how to represent the noise signal $n(t)$. In these calculations it is assumed to be white, that is, uncorrelated across time, making $N(f)$ a constant for all f, denoted ρ. The validity of this assumption will be tested in the experimental section that follows.

The power spectra of $s(t)$ and $n(t)$ are shown in Figure 2. The large low-frequency bias of the training signal $s(t)$ relative to the noise signal $n(t)$ strengthens the hypothesis that a low-pass filter would be effective in removing noise in the training signal. Knowing the form of $S(f)$ and $N(f)$ leads to the following definition of the optimal filter $\Phi(f)$:

$$\Phi(f,\tau,T) = \begin{cases} \frac{\tau^2}{\tau^2+\rho^2T^2} & : \quad f = 0; \\ \frac{4\sin^2\left(\frac{\pi\tau f}{T}\right)}{4\sin^2\left(\frac{\pi\tau f}{T}\right)+\pi^2 f^2\rho^2} & : \quad f \in \mathcal{Z},\ f > 0; \\ 0 & : \quad \text{otherwise}, \end{cases} \tag{2.3}$$

where T represents the period of an entire training epoch, in which all stimuli are seen once, τ corresponds to the period over which all versions of a single stimulus are presented, and Δ is the time for a single stimulus presentation.

Since the training signal is effectively sampled every Δ time units, all signal frequencies above the Nyquist frequency $f_{\mathcal{N}} = 1/2\Delta$ are subject to aliasing. Noise power is hence constrained to lie between zero and $f_{\mathcal{N}}$. Signal amplitude above $f_{\mathcal{N}}$ is assumed negligible, since in general $a_1 \gg a_{f_{\mathcal{N}}}$.[2]

[2] For example, if $T = 100\Delta$ and $\tau = 10\Delta$, $a_1 \approx 50a_{f_{\mathcal{N}}}$.

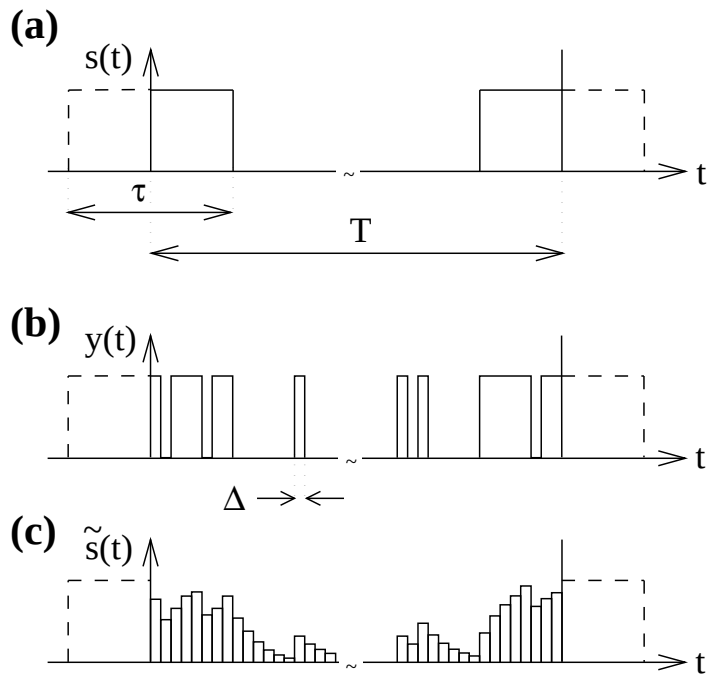

Figure 1: (a) The ideal training signal $s(t)$, which is one when the object is present and zero when absent. (b) Example of a possible neural output signal $y(t)$. (c) Attempt to retrieve $s(t)$ from $y(t)$ by taking a weighted sum of the outputs $y(t)$ at previous times.

The first diagram in Figure 3 depicts the optimal filters derived from the expression for $\Phi(f)$ over a range of noise values,[3] for the example case $\tau = 10\Delta$ and $T = 100\Delta$. Weightings at each time step are proportional to the weighting of the cell's activation at this time in the past, denoted $\bar{\phi}(t)$. The level of noise clearly affects the form of the Wiener filter. In practice, as the neuron learns, its associated error rate will change, and with it, the form of the optimal linear filter.

The large negative trough apparent in all of these filter profiles at time T/τ results from fixing the ratio τ/T. In the next section, this trough is seen to disappear when using variable-length presentation times.

[3] $\rho = 0.45$ is equivalent to an error rate of 10 percent; $\rho = 0.14$ to an error rate of 1 percent.

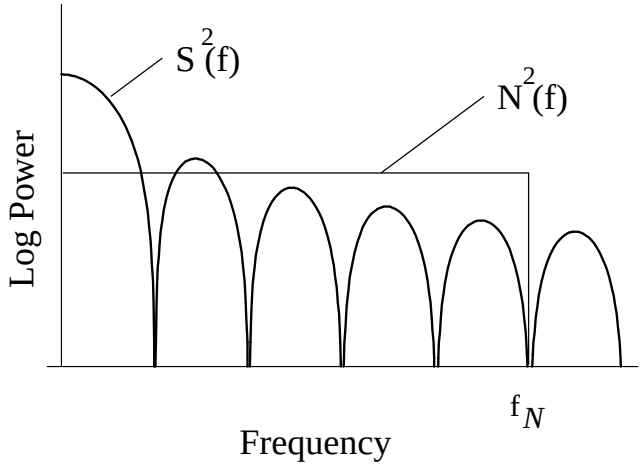

Figure 2: $S(f)$ and $N(f)$ for a system where presentation times are fixed. $f_{\mathcal{N}} = 1/2\Delta$ is the Nyquist frequency.

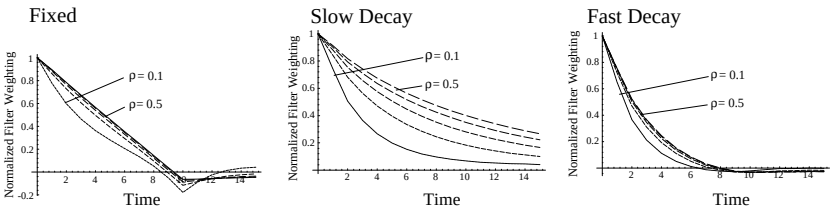

Figure 3: Normalized Wiener filters $\overline{\phi}(t)$ for the three presentation paradigms. The vertical axis represents the weighting accorded to each sample of neural activity; the horizontal axis represents the sample number in multiples of Δ time steps previous to the current sample at time zero.

2.2 Variable-Length Presentation Times. In the real world one might expect to see objects over variable periods of time rather than for the fixed periods. Since presentation time is a scale parameter, it is reasonable to propose that the presentation times are in fact Jeffrey's distributed ($P(\tau) \propto 1/\tau$ or, equivalently, that log presentation times are uniformly distributed (see Jaynes, 1983), which implies that objects tend to be seen for short periods but are occasionally seen for much longer periods.

In this section, new optimal filters are derived from two different presentation probability distributions. In the first, the maximum presentation time for a stimulus, $\hat{\tau}$, is set to be 100Δ, and the minimum presentation length to be Δ. In the second, $\hat{\tau}$ is reduced to 10Δ. The period T is now distributed about a mean dependent on $\hat{\tau}$; hence the mean value of T is given the symbol $T_{\hat{\tau}}$.

Presentation-length probabilities are defined by a simple Jeffrey's distribution extending within the range $\Delta < \tau < \hat{\tau}$, such that $P(\tau)|_{\tau=\hat{\tau}} = 0$. If k is the number of stimulus classes, then we have:

$$P(\tau) = Z_{\hat{\tau}}^{-1}(\tau^{-1} - \hat{\tau}^{-1}) : \Delta \leq \tau \leq \hat{\tau} \tag{2.4}$$

$$Z_{\hat{\tau}} = \sum_{s=\Delta}^{\hat{\tau}} (s^{-1} - \hat{\tau}^{-1}) : \quad \begin{array}{r} Z_{100\Delta} = 4.18738\Delta \\ Z_{10\Delta} = 1.92897\Delta \end{array} \tag{2.5}$$

$$T_{\hat{\tau}} = k \sum_{\tau=\Delta}^{\hat{\tau}} \tau P(\tau) : \quad \begin{array}{r} T_{100\Delta} = 11.8212\Delta k \\ T_{10\Delta} = 2.33285\Delta k\,. \end{array} \tag{2.6}$$

The average optimal filter $\langle \phi(t) \rangle$ is then:

$$\langle \phi(t) \rangle = \sum_{\tau=\Delta}^{\hat{\tau}} P(\tau)\ \phi(t) \tag{2.7}$$

$$= \sum_{\tau=\Delta}^{\hat{\tau}} \left(P(\tau) \sum_{f=0}^{\infty} \Phi(f, \tau, T_{\hat{\tau}}) \cos\left(\frac{2\pi f t}{T_{\hat{\tau}}}\right) \right) . \tag{2.8}$$

The second two graphs of Figure 3 show the filters for these two random presentation distributions. Switching to variable-sequence presentation lengths sees the disappearance of the trough present in the fixed-length presentation case. In addition, the shorter average presentation-time paradigm yields temporal filters with shorter time constants.

2.3 The Trace Rule. This section describes a locally implementable learning rule that can realize the general form of all of the filters described in the previous section. The learning rule produces a running average of neural activity based on a weighted sum of the neuron's previous activity. The simple recursive form of the learning rule may be important in that it lends itself to implementation locally within a cortical neuron (Wallis & Rolls, 1997).

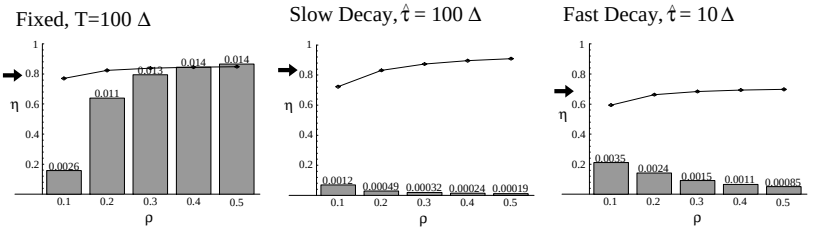

Figure 4: Graphs showing the change in the predicted optimal value of η with changing error rate (black lines and circles). Also shown are the least-mean-squares fit errors at each noise level, shown above the vertical bars. The thick arrows indicate the predicted optimal value of η over the range of noise values shown.

The learning rule has a long history but was used most recently in invariant object recognition for orthogonal images (Földiák, 1991) and nonorthogonal images (Wallis, 1996b), and can be summarized as follows:

$$\Delta w_{ij}{}^{(t)} = \alpha \overline{y}_i{}^{(t)}.x_j \quad : \quad \sum_j w_{ij}{}^2 = 1 \text{ for each } i\text{th neuron} \tag{2.9}$$

$$\overline{y}_i{}^{(t)} = (1-\eta)y_i{}^{(t)} + \eta \overline{y}_i{}^{(t-1)} \tag{2.10}$$

where x_j is the jth input to the neuron, y_i is the output of the ith neuron, w_{ij} is the jth weight on the ith neuron, η governs the relative influence of the trace and the new input, and $\overline{y}_i{}^{(t)}$ represents the value of the ith cell's trace at time t.

The left-most diagram of Figure 4 represents the main result for this section. The optimal value of the trace rule parameter η is plotted as the dark line and varies as a function of the noise ρ. The vertical bars show the relative quality of fit of the filter implemented by the trace rule to the optimal filter as a least-mean-squares fit for the first 20 steps in the filter, with actual fitting errors appearing on the top of each bar. The optimal value of η gradually drops as the noise level drops but remains around 0.8 over a very wide range of noise, suggesting that a constant trace rule could be used throughout training.

The remaining two graphs of Figure 4 show the change in the best-fitting value of η as a function of ρ, for the variable presentation times. The graphs are similar to the fixed presentation time case, except that the optimal value of η is a little higher for the slow-decaying case (between 0.8 and 0.9) and somewhat lower for the fast-decaying case (0.7). The average fitting error is also considerably smaller, indicating that the training rule fits the optimal filter more closely under the two probabilistic presentation regimes. The

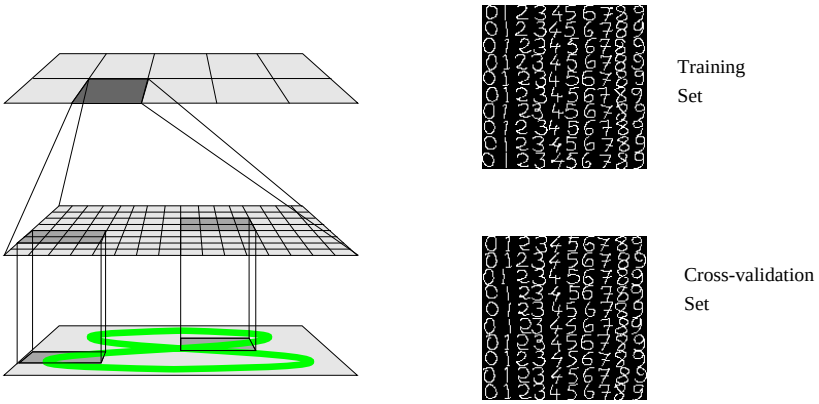

Figure 5: Architecture of the network used in the simulations and the two sets of digits used during training and testing.

accuracy of these and previous predicted optimal values of η is tested in the next section.

3 Simulating the Predictions of the Wiener Filtering

A series of predicted optimal linear filters, parameterized for stimulus presentation length and the classification error rate of the neurons ρ, have now been produced. To gauge the accuracy of the predictions, a series of simulations were run.

3.1 Methods. A two-layer network was constructed (see Figure 5). The first layer acts as a local feature extraction layer and consists of a 16×16 grid of neurons arranged in 4×4 pools. Each pool fully samples a corresponding 4x4 patch of a 16×16 input image. All learning in this layer is standard Hebbian. Above the input layer is a second layer, consisting of a single inhibitory pool of 10 neurons, which fully samples the first layer. Neurons in this layer are trained with the trace rule. Competition acts within each pool in both layers and is implemented using the soft max algorithm (Bridle, 1990).

Digits from a character set used by LeCun et al. (1989) were presented to the network. During learning, 100 digits—10 of each type—were presented in permuted random sequence according to the three stimulus presentation distributions described in the previous section. Such a stream of digits might

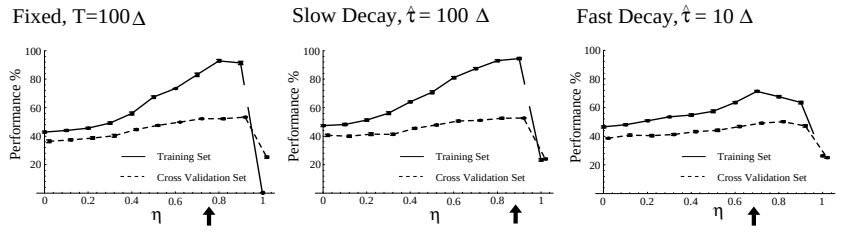

Figure 6: Classification performance for the network at differing values of η for the three presentation-length paradigms. The thick arrows indicate the optimal value of η predicted by the earlier theory, which are seen to be in accord with the optimal values to come out of the simulations.

be generated in the real world while observing a digit printed in a book, by altering the reading distance or tilting a page.

Although we have chosen to concentrate on learning deformation invariance of digits here, this system has been shown to be able to cope with other transformations of more natural stimuli, such as faces, undergoing depth rotations, scale changes, and translations (Wallis & Rolls, 1997; Wallis, 1996a).

Network performance was measured as both recognition of the 100 digits from the training set and a 100 digit cross-validation set from the same database. In addition to these two performance measures, the amount of noise present in a neuron's output signal when trained from start to steady response was also recorded.

3.2 Results. The results shown in Figure 6 reveal the effect on overall classification performance of changing the value of the trace parameter η over 10 runs of the network. Under each of the three presentation paradigms, the optimal value of η predicted in the previous section is seen to correspond very closely to the peak in the performance graph displayed alongside (see the dark arrows in each of the six graphs). Therefore, despite two possible theoretical problems (the assumptions of independent errors and of linearity), the predictions from the theory are very close to the results of the simulations.

For $\eta = 0$ the results correspond to simple Hebbian learning. These results are clearly much worse than those achieved using the optimized trace rule. Further comparisons with other architectures and other learning rules are described elsewhere (Wallis, 1996b).

The forerunning calculations depend on the errors' being approximately random. Figure 7 shows the average power spectrum recorded for a neuron

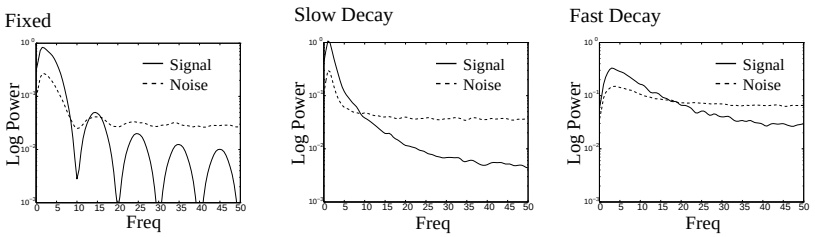

Figure 7: Average log power spectrum for noise $n(t)$ and ideal training signal $s(t)$ in the response of neurons under each of the three presentation-length training paradigms.

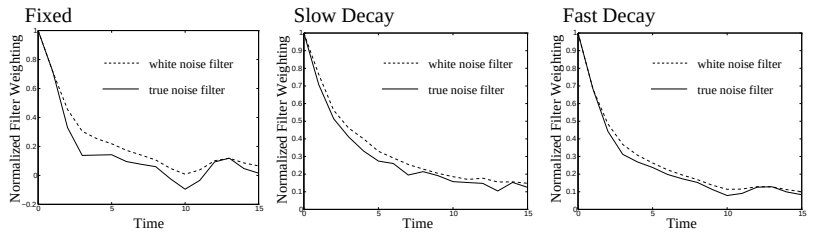

Figure 8: Optimal filters calculated from the measured neural noise under each of the three presentation-length training paradigms. The noise signal used to calculate the filters was assumed to be either "white," that is, uniform across all frequencies (dashed line), or the true low-frequency biased signal (solid line).

monitored from a random starting point to the point at which a steady-state response had been achieved. The spectrum is averaged across all 10 output cells and also across time, with the signals being binned into 100 consecutive stimulus presentations. All three graphs show a peak in the low-frequency end of the noise spectra, although the power of this signal is considerably less than the signal power, and the spectrum is essentially flat for all higher frequencies up to and beyond the point at which noise power exceeds signal power, in accordance with the form of the graph in Figure 2.

Knowing the true noise signal permits recalculation of the optimal filters. Figure 8 shows the optimal filters obtained using the true low-frequency biased noise spectrum versus using a flat (white) spectrum, taken as the average of the true noise power across all frequencies.

The optimal filter for the true noise case is very similar to the white noise case, although the decay through time is consistently slightly faster. The

fact that this difference is consistent may well be attributed to the fact that one effect of the peak at low frequencies in the true noise signal is to shift the frequency at which the noise and signal spectra cross in comparison with the crossing point for the flat, white noise case. The white noise curve would cross the signal curve at the same frequency as in the true noise case if its power was reduced slightly, and so the change from white to true noise can be thought of as a reduction of the effective white noise amplitude. Figure 4 showed how lower noise values correspond to smaller time constants, consistent with the shift seen in Figure 8.

Despite this shift, the stability of the optimal value of η with respect to changes in noise (see Figure 4) ensures that the predicted optimal value of η never differs by more then 0.05 between the white and true noise cases, supporting the use of a flat noise spectrum in the theory section of this article.

4 Conclusions

This article has proposed that objects encountered in the world are presented in sequences of views of the same object, not randomly. This temporal regularity can then be used to help form invariant representations by training on the basis of a signal generated from a weighted sum of previous neural activity, the form of which can be determined by optimal filtering theory. Further, if we make the reasonable assumption that the log of presentation times is uniformly distributed, then this weighting can be achieved by a simple local learning rule.

The validity of this theory has been tested on the recognition of digits, demonstrating that optimal parameters were well predicted and that assumptions about the form of the noise and system linearity were justified.

In cases where we have explicit labels for inputs, supervised learning techniques may well be preferable, but in the natural world, the visual input to the brain often does not have such labels. In this case, any regularity in the input (spatial or temporal) should be exploited by an unsupervised system to learn useful representations. Much work has concentrated on exploiting spatial correlations. Here we show that temporal regularities can also be exploited near optimally using the simple and local trace rule.

Acknowledgments

We are grateful to two anonymous reviewers and this journal's editor in chief for helping to improve the clarity and depth of this article considerably.

References

Bridle, J. S. (1990). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters.

In D. S. Touretzky (Ed.), *Neural Information Processing* (Vol. 2, pp. 211–217). San Mateo, CA: Morgan Kaufmann.

Bülthoff, H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Science, USA*, *92*, 60–64.

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

Edelman, S., & Weinshall, D. (1991). A self-organising multiple-view representation of 3D objects. *Biological Cybernetics*, *64*, 209–219.

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

Jaynes, E. T. (1983). *Papers on probability, statistics and statistical physics*. Dordrecht: Reidel.

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

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

Perrett, D., & Oram, M. W. (1993). Neurophysiology of shape processing. *Image and Vision Computing*, *11*(6), 317–333.

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

Press, W. H. (1992). *Numerical recipes in C: The art of scientific computing*. Cambridge: Cambridge University Press.

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

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

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psycholgy*, *21*, 233–282.

Wallis, G. (1996a). Temporal order in object recognition learning. To appear in *Journal of Biological Systems*.

Wallis, G. (1996b). Using spatio-temporal corrolations to learn invariant object recognition. *Neural Networks*, *9*, 1513–1519.

Wallis, G., & Rolls, E. T. (1997). A model of invariant object recognition in the visual system. *Progress in Neurobiology*, *51*: 167–194.

Wiener, N. (1949). *Extrapolation, interpolation, and smoothing of stationary time series: With engineering applications*. New York: Wiley.

Received February 8, 1996; accepted September 4, 1996.