LETTER Communicated by Jack Cowan

A Neural Network Model of Temporal Code Generation and Position-Invariant Pattern Recognition

Dean V. Buonomano
Michael Merzenich
Keck Center for Integrative Neuroscience, University of California, San Francisco, San Francisco, CA 94143, U.S.A.

Numerous studies have suggested that the brain may encode information in the temporal firing pattern of neurons. However, little is known regarding how information may come to be temporally encoded and about the potential computational advantages of temporal coding. Here, it is shown that local inhibition may underlie the temporal encoding of spatial images. As a result of inhibition, the response of a given cell can be significantly modulated by stimulus features outside its own receptive field. Feedforward and lateral inhibition can modulate both the firing rate and temporal features, such as latency. In this article, it is shown that a simple neural network model can use local inhibition to generate temporal codes of handwritten numbers. The temporal encoding of a spatial patterns has the interesting and computationally beneficial feature of exhibiting position invariance. This work demonstrates a manner by which the nervous system may generate temporal codes and shows that temporal encoding can be used to create position-invariant codes.

1 Introduction

Experimental (McClurkin, Optican, Richmond, & Gawne,1991; Middlebrooks, Clock, Xu, & Green, 1994; Bialek, Rieke, Ruyter van Steveninck, & Warland, 1991) and theoretical (Hopfield, 1995; Thorpe & Gautrais, 1997) studies have suggested that the nervous system may encode information in the temporal structure of neuron spike trains, generally referred to as temporal coding. For example, McClurkin et al. (1991) have shown that by taking into account the temporal structure of neuronal responses to Walsh patterns, there is more information about the stimuli than there is in the firing rate alone. However, in addition to showing that there is information in the temporal structure of spike trains, at least two additional issues relating to temporal encoding must be addressed: (1) How does sensory information come to be temporally encoded? (2) How is temporally coded information used or "decoded" by the nervous system? Here we focus on the first question and consider the potential advantages of encoding information in the temporal domain.

In principle, any of the many neuronal properties that affect the balance of excitation and inhibition can produce significant changes in the temporal structure of neuronal responses. If the temporal structure is to contain information about a given stimulus, it should be reproducible and stimulus specific, and should be amenable to stimulus generalization. Local inhibitory interactions, which include both feedforward and lateral inhibition, may provide a neural mechanism that satisfies these conditions. Local inhibition is an almost ubiquitous feature of neuronal circuits and typically has been thought to be a means of preventing cells from firing or of modulating the average firing rate of neurons. However, inhibition can also shape the temporal structure of responses by modulating when a neuron will reach threshold. As a result of local inhibition, both the firing rate and temporal structure of neuronal responses can be significantly altered by neighboring neurons and may contain information not only about their own receptive fields but also about the local spatial structure of stimuli. Changes in the temporal structure of neuronal responses can be manifested in various degrees of complexity. The simplest feature that may be changed, from both an encoding and decoding perspective, is spike latency, which we will consider here.

One possible advantage of temporal encoding is that the amount of information that can be potentially transmitted on a per-spike basis is larger than that transmitted by a rate code. An alternate or additional advantage of temporal coding is that once spatial information is temporally encoded, it can potentially represent spatial patterns in a position-invariant manner. Position invariance has proved to be an extremely challenging problem, from the perspective of understanding how the brain solves it and in developing artificial systems capable of invariant pattern recognition. In their simplest forms, conventional neural networks do not exhibit position invariance because information is stored in the spatial pattern of synaptic strengths. Figure 1 provides a schematic illustration of why conventional neural networks are not generally well suited to solve position invariance. If a unit is to behave as a + detector, the connection strengths of the weight matrix are distributed in a fashion that spatially reflects the + symbol. If the + is shifted to different positions on the input layer, the + detector may develop responses to other stimuli. To circumvent this problem, various biologically plausible models have been proposed that are capable of exhibiting different forms of invariant pattern recognition. One approach has been to develop large-scale multilayer networks, in which each layer exhibits position invariance to higher-order features (Fukushima, 1988). A second approach has been to develop networks capable of dynamically changing the local connectivity (Olshausen, Anderson, & Van Essen, 1993; Konen, Maurer & von der Malsburg, 1994), or the local gain of the network (Salinas & Abbott, 1997), thus essentially accomplishing online translation and scaling of images. Here we develop an alternative hypothesis based on temporal coding.

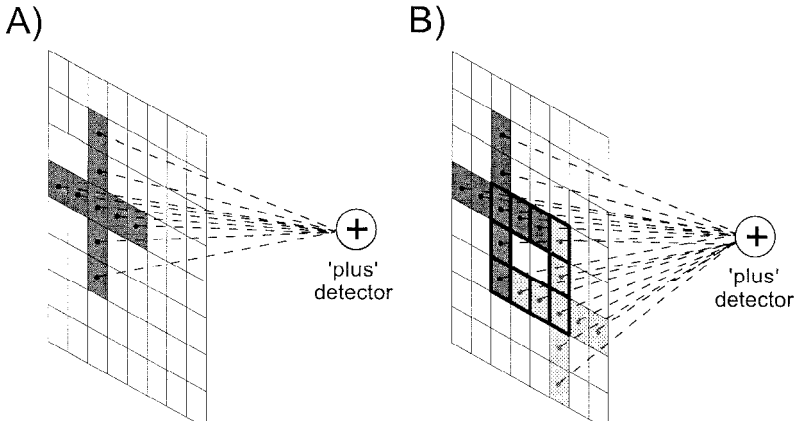

Figure 1: Schematic of the position-invariant pattern recognition problem. (A) In a conventional one-layer network, the creation of a "plus" detector essentially consists of connecting the spatial arrangement of units activated by the + to a plus-detector output unit. (B) If the same plus detector is also going to detect a + in a different position, such as the lower left corner, the units activated by that + will also be connected to the plus detector. In the process, other patterns, such as a square (bold outline), will also activate the plus detector.

2 Methods

Figure 2 schematizes the type of feedforward, surround inhibition used in the model. Figure 3 shows a simple network that captures the basic principles of our model. Below we explain in detail the more complex network used for handwritten digit recognition.

Each unit from the input layer (the "retina") provides an excitatory input to the topologically equivalent position in the feature detector layer (analogous to L-IV of V1), and inhibition to the neighboring units in a surround inhibition pattern. Each point in the feature detector layer had four types of line detectors (vertical, horizontal, and two diagonals). Each feature detector unit was activated if five appropriately aligned input units were on. The feature detector layer provided input to the "cortical" layer (analogous to L-II/III of V1). The voltage (V_i^f) of a cortical unit in position i, and of

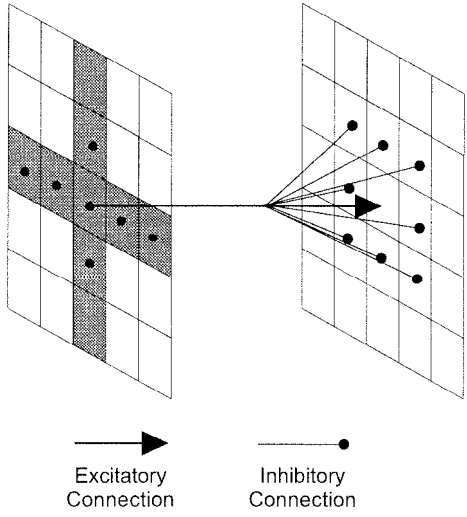

Figure 2: Schematic of the circuit used for feedforward, center surround inhibition. Units from the input layer provide excitatory input to the topologically equivalent unit in the next layer and inhibitory connections to the neighboring units.

orientation f, is given by

$$V_i^f(t) = V_i^f(t-1) + K_{Ex} I_i^f - \sum_{l \neq f}^{F} \sum_{j}^{INH} I_j^l \cdot W_{ij} + \sum_{k \neq i}^{EX} I_k^f \cdot W_{ik} \,. \tag{2.1}$$

I_i^f represents the binary input from the unit in position i and of orientation f from the feature detector layer. The third term represents the inhibition, which is implemented in a cross-orientation manner; for example, each vertical unit receives inhibition from the horizontal and two diagonal input layers. The weights, W_{ij}, are a function of the distance from unit i and of the difference in orientation preference between the units:

$$W_{ij} = K_F \cdot \exp(-\|i - j\|) \cdot K_{Inh} \,, \tag{2.2}$$

where $K_F = 1$ when the orientation of i and j differs by 90 degrees and $K_F = 0.66$ when they differ by 45 degrees. The fourth term in equation 2.1 represents iso-orientation excitation; a vertical unit will excite neighboring

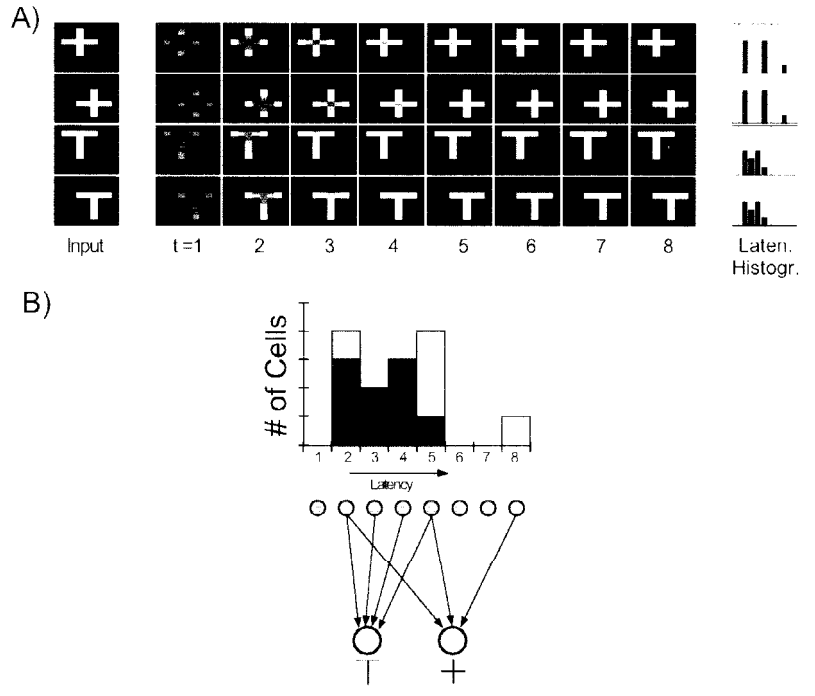

Figure 3: A simple retinal model with inhibition can create position-invariant temporal codes. (A) The first column of each row represents the input pattern—either a + or a T at two different positions. The subsequent columns represent frames of the voltage of each element in time. Voltage is represented on gray scale. A unit in white represents those that have reached threshold. Since we are interested in onset latency, what happens after a unit reaches threshold is not relevant, and it is assumed that each element stays on. The time step at which threshold was reached defines the latency of that unit. The latency histogram produced by each image is represented in the last column. The *y*-axis represents the number of activated units at a given latency. (B) Schematic diagram of a network that could use latency histograms for position-invariant pattern recognition (empty and filled bars represent + and T, respectively). The latency code is used as spatial code in which each unit corresponds to a latency and serves as the input to the recognition network.

vertical units that are vertically aligned, but not vertical units that are horizontally aligned with it. For excitatory weights $W_{ij} = C \cdot 0.04$, where C is 1.0, 0.67, or 0.0, when j is positioned above or below, diagonal to, or lateral to

i, respectively, for a vertical unit. Such iso-orientation topography has been reported experimentally (Fitzpatrick, 1996). Threshold was arbitrarily set at 1, and $V(0) = 0$.

One important parameter is the radius of the of the feedforward inhibition. If the radius is too small in relation to the size of the image, all local interactions are likely to be expressing local noise rather than the relevant structure. Conversely, a large radius will tend to "normalize" all responses according to overall activity, and not express any local structure. In the current simulations, an inhibition radius of 3 was found to be optimal. Within a fairly wide range, changes in the excitation K_{Ex} and inhibition K_{Inh} constants did not dramatically affect performance as long as inhibition was strong enough to modulate the latency. In the simulations presented here, the values of K_{Ex} and K_{Inh} were 0.26 and 0.002, respectively.

2.1 Recognition Network. The second component of the model consisted of a recognition network, which was necessary in order to determine whether the temporal codes generated could actually be used for position-invariant pattern recognition. For this purpose, it is necessary to decode the temporal code. Decoding temporal codes is critical not only when the nervous system may use internally generated temporal codes, but in temporal processing in general, an important and difficult problem faced by the brain and artificial networks processing time-varying signals, such as speech. Presenting a realistic model of temporal processing is far beyond the scope of this article. Indeed, it is because the decoding of complex temporal codes can easily become intractable that we chose to analyze latency of the first spike.

To decode temporal information—that is, transform a temporal code into a spatial code—we assumed the presence of delay lines (e.g., Tank & Hopfield, 1987; Waibel, 1989), in which hypothetical delays are used to generate elements sensitive to temporal features, specifically latency. In practice, each of the latency histograms represents the output of the delay line model. Each bin represents the output of a single element tuned to a given delay (schematized in Figure 3b).

The next step was to determine whether once the latency codes are mapped spatially, they can be used for digit recognition. For this purpose, the output of the delay line network was used in a conventional backpropagation network (Rumelhart & McClelland, 1986). Since each orientation sublayer generates a temporal code, and only 10 different latencies were considered, the input to the backpropagation network consisted of a single vector of length 40. The backpropagation network contained 16 hidden units and 10 output units (one for each numeral).

The handwritten digit database was obtained from the National Institute of Standards and Technology. The samples used here were from the Handprinted Character Database, subdirectory f13/fl3/data/f0035, which

comprise characters from a single writer. The raw images were used; that is, no scaling or normalization was performed.

3 Results

Figure 3 illustrates a simple network that captures the fundamental properties of our model, in which temporal codes of images are generated with the use of local inhibition. These codes can be then used by a second network for position-invariant pattern recognition. In this simplified network, a single-layer network of linear-threshold units was used. Each "pixel" of the input stimulus excites a unit in the corresponding position of the network and inhibits neighboring units (see Figure 2). During each time step, the sum of excitation and inhibition was computed. If the sum reaches threshold, a spike occurs. In the simulations considered here, only the first spike is relevant; once a unit has fired, it remains on. The latency is defined as the time step at which threshold was reached. In the simulation shown in Figure 3A, two symbols (T and +), each composed of a single vertical and horizontal line, were presented to the network, each at two different locations. For the +, the four extremes reached threshold first and fired at $t = 2$. These four units fired first because each receives inhibition from only one inhibitory input. The center unit will be the last to fire at $t = 8$, because it is inhibited by four neighboring units. In contrast for the T, three extreme units will fire first, followed by their neighbors, until the innermost unit fires. The plots on the right of Figure 3A show the latency histograms of the network. Note that the latency histograms are distinct for the T and + and are independent of position. Figure 3B illustrates how the latency distributions could be used for position-invariant pattern recognition. Using a traditional neural network architecture, the latency histograms serve as the input patterns, and each output unit becomes a pattern detector. An implicit step in this process is to transform the temporal code into a spatial code. For example, this can be accomplished using tapped delay lines.

To determine whether neural networks that generate temporal codes can be used to recognize real-world patterns, we developed a network and tested it with handwritten digits placed in different locations on the input layer (see Figure 4). The network consisted of an input layer, a feature detection layer, and a "cortical" layer. The feature detection and cortical layer each contained a complete topographic representation of four different orientations. The feature detection and cortical layer can be best visualized as each containing four distinct sublayers, one for each orientation: vertical, horizontal, and two diagonal sublayers. Lateral interactions take place in the cortical layer in the form of cross-orientation inhibition and iso-orientation excitation.

Figure 5 shows the result of a simulation. The input pattern activates the appropriate units on the vertical, horizontal, and diagonal feature detector layer. The feature detector layer projects to the cortical layer, in which the

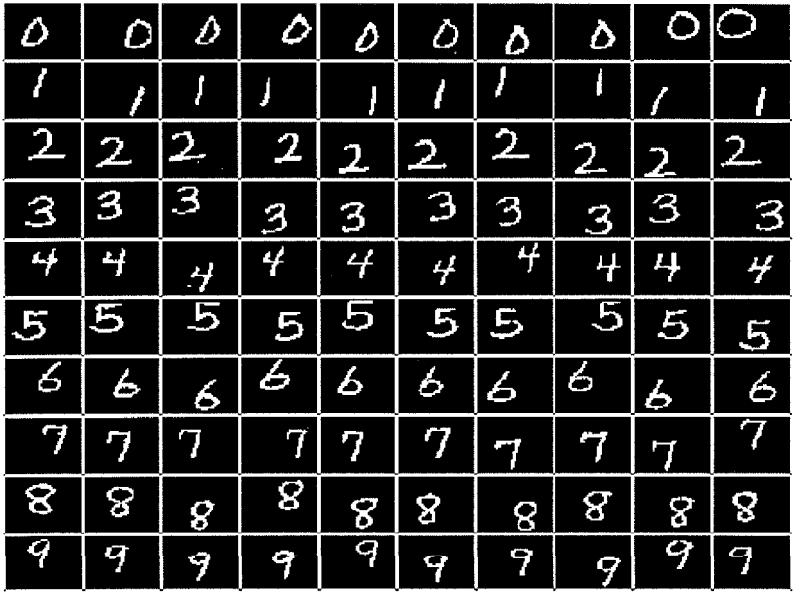

Figure 4: Handwritten digits used to test position-invariant pattern recognition. Digits were obtained from a National Institute of Standards and Technology database in a 32 × 32 pixel format. Each digit was placed at random location on a 64 × 64 input layer in order to test for position invariance.

total inhibitory and excitatory input is computed for each unit on each time step. For visualization purposes, the cells of the same orientation preference are shown as a separate sublayer. This is similar to looking down on V1 and looking at all horizontal, vertical, and diagonal cells separately. Each of the four cortical sublayers generates a latency histogram, which represents how many units fired at each time step (only the first spike is relevant). To determine whether these temporal codes latency histograms are sufficient to code for all digits in a position-independent manner and if it could generalize across samples, we presented 100 handwritten digits to the network (see Figure 4). Each 32 × 32 digit image was placed at a random location on the 64 × 64 input layer. Each of the 100 sets of four latency histograms served as an input vector to the recognition network, which was a standard backpropagation network. Figure 6A (upper panel) shows an example of the 50 latency histograms used for testing. The latency histograms for each of the sublayers are placed in a single row. The recognition network was

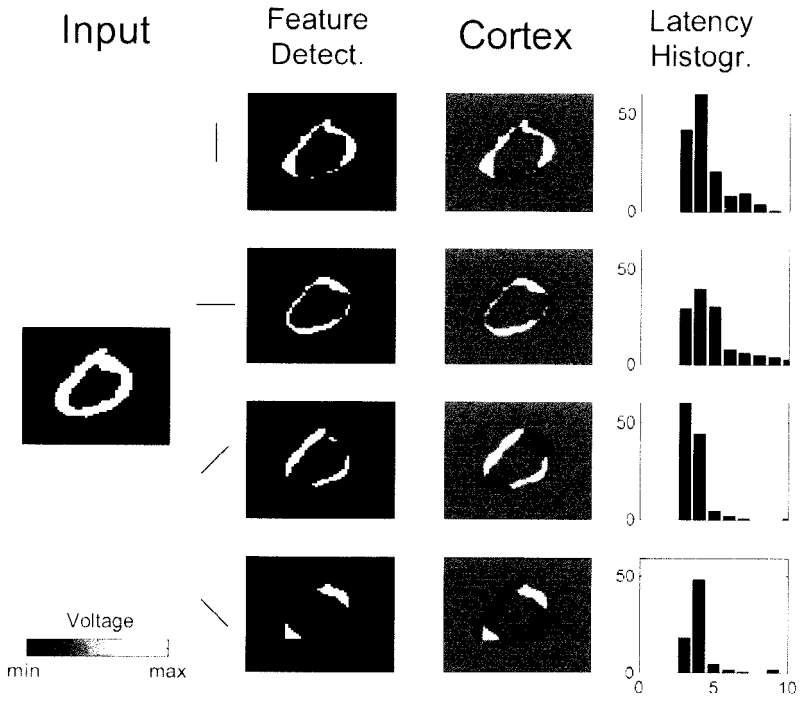

Figure 5: Simulation of digit recognition. An image is presented to the input layer, which is preprocessed by a feature detector layer. The feature detector layer has four types of orientation detectors: vertical, horizontal, and two diagonal lines. For visualization, each type of detector is presented as a separate sublayer. The feature detector layer then projects to a "cortical" layer, which also contains a representation of each orientation. Units in the feature detector layer inhibit units in the cortical layer in a cross-orientation fashion and excite them in an iso-orientation fashion. The voltage in the cortical units is shown at $t = 10$. When and how many cells reached threshold at a given time step is shown in the latency histograms on the right.

trained on 50 latency histograms (5 of each digit), and then tested on the remaining 50 patterns. The lower panel in Figure 6A shows the output of the recognition network in response to the 50 test latency histograms. In this run, 48 of 50 digits were correctly classified. The average performance was $93.4 \pm 0.16\%$.

It is important to determine that the performance of the network is de-

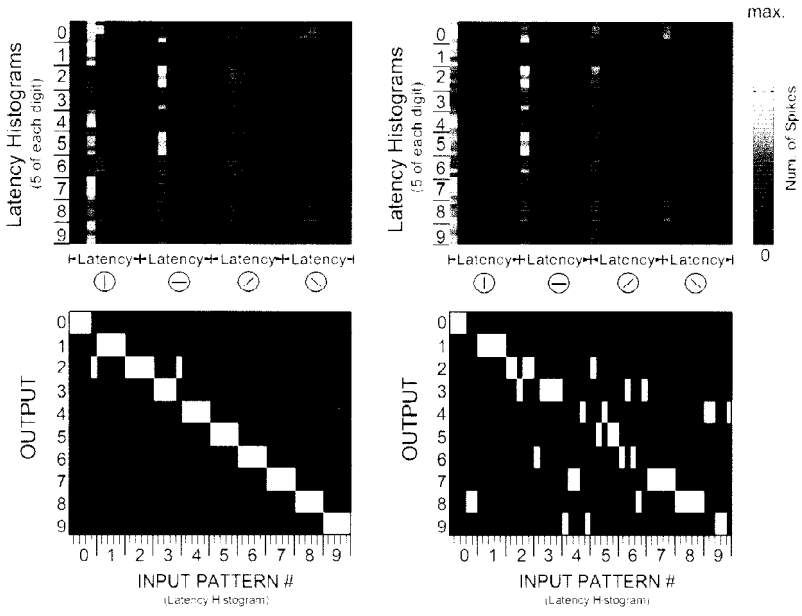

Figure 6: (A) Latency histograms for the 50 digits used to test network performance represented in a gray-scale code (upper panel). A black point means that no units fired at that latency in response to a given digit; white means that a maximal number of units fired at that latency. The lower panel represents and the output of the backpropagation network in response to the 50 test digits, after training on the 50-digit learning set. (B) Control histograms in which temporal information is collapsed (upper panel) and output of the backpropagation network trained on the control stimulus set.

pendent on temporal information rather than on spatial information. Since there are four different types of feature detectors, there is information contained in the total number of spikes from each feature detector sublayer irrespective of temporal structure. For example, since the number 1 essentially corresponds to the vertical feature detector, it is unlikely that temporal information is contributing to its recognition. To examine the contribution of the latency code, temporal information was removed by collapsing each of the four latency histograms into a single time bin and training the same recognition network on the collapsed input vectors. The upper panel in Figure 6B shows the input patterns for the 50 test stimuli when the latency histograms are collapsed across each of the four cortical sublayers. The lower panel in Figure 6B shows the output of the pattern recognition networks.

In this example, 33 of 50 of the digits were correctly classified. Note that, as expected, the exemplars of digit 1 were well classified based solely on spatial information. On average, performance was below 70% in the absence of temporal information.

In neural network models, it is generally important to determine how noise affects the performance of the network. Both extrinsic and intrinsic noise sources should be considered. Extrinsic noise refers to the noise or variability of the stimulus set. Since we have used a real-world stimulus set, the model presented here clearly performs well with the inherent variability generated by a single writer (see Figure 4). To provide some insight as to the performance of the network in the presence of intrinsic noise, we assumed that all elements of the cortical network exhibited some level of spontaneous activity. Assuming the time steps in our simulations correspond to approximately 1 ms, we examined performance in the presence of 1 Hz and 10 Hz spontaneous activity. At 1 Hz noise, there was a small drop in the performance to $87.44 \pm 0.36\%$. At 10 Hz, there was a large drop in performance to $60.3 \pm 0.77\%$. We should stress that it is difficult to compare the effect of noise on simple artificial networks with that of biological networks. Sources of intrinsic noise have various biological sources, including synaptic and membrane potential variability. Even when these data are available, it is difficult to apply them to simple models such as that presented here in which arbitrary discrete time steps are used. Additionally, there is generally a trade-off between noise levels and the size of the network; thus, the behavior of small networks is generally more sensitive to noise.

4 Discussion

The results described here show that by using temporal information generated by local inhibition, it was possible to create a network that classified handwritten digits in a position-invariant fashion. The temporal codes generated for each pattern were used to train the recognition network; half the stimuli were used for training and half for testing. After training, the network generalized appropriately to the test stimuli, comprising both different digit exemplars and positions. The ability to recognize different handwritten exemplars indicates that the temporal codes are sufficiently specific to code for the 10 different digits, yet robust enough to generalize over the intrinsic variability of the digits. Our stimulus set was from a single writer; stimulus sets from multiple writers will decrease the performance of the simple network presented here.

Good performance of the network was obtained despite a simple implementation; specifically, only four feature detectors were used. There are a few intrinsic limitations of using simple feature detectors with 180-degree symmetry. For example, the network cannot distinguish between the same image rotated by 180 degrees, since precisely the same units will be activated. Nevertheless, the network correctly classified most instances of 6

and 9 because of distinct local features intrinsic to the handwritten samples. (See the upper left panel of Figure 6. More vertical elements are activated by 9. Additionally, most are activated on time step 3, whereas more vertical elements are activated at time step 4 for the 6.) More sophisticated implementations of the model could be custom designed for specific pattern recognition problems by incorporating more and higher-order sets of feature detectors.

It is the presence of local interactions in general, rather than the specific characteristics of the connectivity, that underlies the generation of temporal codes. In other words, the model does not rely on the specific assumptions of cross-orientation inhibition and iso-orientation excitation.

We propose that one of the functions of local interactions in neural circuits may be to generate temporal codes. Such local circuit interactions would represent a simple, widespread, and biologically plausible mechanism by which the nervous system could encode information by extending it into the temporal domain. Indeed, the temporal structure of neurons has been reported to contain a significant amount of information (McClurkin et al., 1991; Middlebrooks et al., 1994; Bialek et al., 1991). In the current model, we have simplified the problem of both generating and analyzing temporal codes by focusing on the latency of the first spike (see also Thorpe and Gautrais, 1997). We suspect that the same circuitry will generate more complex temporal codes that are likely to increase the performance of the network and make it less sensitive to intrinsic noise; however, more enhance the richness of the temporal code, but would require a more sophisticated decoding stages would also likely become necessary. Furthermore, analysis of latency is reasonable since it may be one of the most important temporal parameters (Richmond, Optican, & Spitzer, 1990; Tovée, Rolls, Treves, & Bellis, 1993; Gawne, Kjaer, & Richmond, 1996).

Here we have shown that temporal coding may emerge from simple and well-established network characteristics. Furthermore, we have suggested that position invariance may be a reason for which it is beneficial to encode information in the temporal domain. Our model, which relies on local circuit interactions, establishes a biologically plausible mechanism to generate temporal codes that have been proposed to contribute to information processing (Hopfield, 1995). Thorpe and Gautrais (1997) have also proposed that different spike latencies, in their model generated as a function of contrast, could be used to generate temporal codes that could be used for pattern recognition.

We should emphasize two potential limitations of the hypothesis presented here. First, the mechanisms proposed here cannot be solely responsible for position invariance; clearly, position information is not collapsed across large portions of retinal position in the early stages of visual processing. However, the temporal codes generated by lateral interactions could to contribute to position invariance, particularly on small scales in a multistage process. Second, other stimulus features, such as contrast, that modu-

late spike latency could easily confound temporal codes, although temporal codes generated at higher levels of visual processing or normalization mechanisms could be used to overcome this problem. One of the predictions that emerges from our hypothesis is that the response latency or temporal structure of neurons to simple images such as + and T (see Figure 1) should be different. Since the differences in temporal structure arise from feedforward and lateral interactions, they are likely to be more robust in higher-order versus primary visual areas.

In addition to the problem of temporal encoding, a critical issue that remains to be addressed, if the nervous system uses temporal codes, is that of decoding. We did not address this issue here, and a simple version of delay lines was used in decoding the temporal patterns. However, delay lines are unlikely to account for temporal decoding, and temporal processing in general, particularly for more complex temporal patterns. Networks that rely on local circuit dynamics and short-term forms of plasticity may provide a more biological mechanism to decode temporal information (e.g., Buonomano & Merzenich, 1995), but future experimental and theoretical research must focus on how the brain decodes temporal information, in addition to temporal encoding of information.

Acknowledgments

The work was supported by ONR grant N00014-96-1-0206. We thank C. deCharms, M. Kvale, H. Mahncke, Jennifer Raymond, F. Theunissen, and members of the Lisberger lab for helpful discussions and comments of an earlier version of this article.

References

Bialek, W., Rieke, F., Ruyter van Steveninck, R. R., & Warland, D. (1991). Reading a neural code. *Science, 252*, 1854–1857.

Buonomano, D. V., & Merzenich, M. M. (1995). Temporal information transformed into a spatial code by a neural network with realistic properties. *Science, 267*, 1028–1030.

Fitzpatrick, D. (1996). The functional organization of local circuits in visual cortex—Insights from the study of tree shrew striate cortex. *Cerebral Cortex, 6*, 329–341.

Fukushima, K. (1998). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks, 1*, 119–130.

Gawne, T. J., Kjaer, T. W., & Richmond, B. J. (1996). Latency—Another potential code for feature binding in striate cortex. *J. Neurophysiol., 76*, 1356–1360.

Hopfield, J. J. (1995). Pattern recognition computation using action potential timing for stimulus representation. *Nature, 376*, 33–36.

Konen, W. K., Maurer T., & von der Malsburg, C. (1994). A fast dynamic link matching algorithm for invariant pattern recognition. *Neural Networks, 7*, 1019–1030.

McClurkin, J. W., Optican, L. M., Richmond, B. J., & Gawne, T. J. (1991). Concurrent processing and complexity of temporally encoded neuronal messages in visual perception. *Science, 253*, 675–677.

Middlebrooks, J. C., Clock, A. E., Xu, L., & Green, D. M. (1994). A panoramic code for sound location by cortical neurons. *Science, 264*, 842–844.

Olshausen, B. A., Anderson, C. H., & Van Essen, D. V. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci., 13*, 4700–4719.

Richmond, B. J., Optican, L. M., & Sptizer, H. (1990). Temporal encoding of two-dimensional patterns by single units in primate visual cortex. I. Stimulus-response relations. *J. Neurophysiol., 64*, 351–368.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing*. Cambridge, MA: MIT Press.

Salinas, E., & Abbott L. F. (1997). Invariant responses from attention gain fields. *J. Neurophysiol., 77*, 3267–3272.

Tank, D. W., & Hopfield, J. J. (1987). Neural computation by concentrating information in time. *Proc. Natl. Acad. Sci., 84*, 1896–1900.

Thorpe, S. J., & Gautrais, J. (1997). Rapid visual processing using spike asynchrony. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.), *Advances in neural information processing systems, 9* (pp. 901–907). Cambridge, MA: MIT Press.

Tovée, M. J., Rolls, E. T., Treves, A., & Bellis, R. P. (1993). Information encoding and the responses of single neurons in the primate temporal visual cortex. *J. Neurophysiol., 70*, 640–654.

Waibel, A. (1989). Modular construction of time-delay neural networks for speech recognition. *Neural Comp., 1*, 39–46.

Received October 10, 1997; accepted May 7, 1998.