Using Spatiotemporal Correlations to Learn Topographic Maps for Invariant Object Recognition

Frank Michler, Reinhard Eckhorn, and Thomas Wachtler
NeuroPhysics Group, Philipps-University Marburg, Marburg, Germany

Submitted 9 June 2008; accepted in final form 29 May 2009

Michler F, Eckhorn R, Wachtler T. Using spatiotemporal correlations to learn topographic maps for invariant object recognition. J Neurophysiol 102: 953–964, 2009. First published June 3, 2009; doi:10.1152/jn.90651.2008. The retinal image of visual objects can vary drastically with changes of viewing angle. Nevertheless, our visual system is capable of recognizing objects fairly invariant of viewing angle. Under natural viewing conditions, different views of the same object tend to occur in temporal proximity, thereby generating temporal correlations in the sequence of retinal images. Such spatial and temporal stimulus correlations can be exploited for learning invariant representations. We propose a biologically plausible mechanism that implements this learning strategy using the principle of self-organizing maps. We developed a network of spiking neurons that uses spatiotemporal correlations in the inputs to map different views of objects onto a topographic representation. After learning, different views of the same object are represented in a connected neighborhood of neurons. Model neurons of a higher processing area that receive unspecific input from a local neighborhood in the map show view-invariant selectivities for visual objects. The findings suggest a functional relevance of cortical topographic maps.

INTRODUCTION

Invariant object recognition

Our visual system has the capability of invariant object recognition: we recognize a familiar object under different viewing conditions, despite drastic variations in the corresponding retinal images with viewing angle, distance, or illumination. Physiological studies have shown that cells in monkey V4 and inferotemporal cortex (Ito et al. 1995; Tanaka 1996, 2003; Tovee et al. 1994; Wang et al. 1996) and in the human hippocampus (Quiroga et al. 2005) show selectivity for objects invariant of size or viewing angle.

A prototype for models of invariant representations is the pooling model (Hubel and Wiesel 1962; Kupper and Eckhorn 2002; Riesenhuber and Poggio 1999). An output cell receives input from a pool of cells that have the same selectivity in one feature dimension, but a different selectivity in a second feature dimension. The output cell will then respond selectively to the first feature, but will show invariant responses with respect to the second feature.

Spatial and temporal stimulus correlations as cues for learning invariant representations

When we move through our environment while fixating an object, or when we manipulate an object, different views of the same object appear in temporal sequence. The retinal projections change continuously, whereas the identity of the object remains the same. Under such natural viewing conditions, projections of different views of the same object are spatially and temporally correlated. Physiological (Miyashita 1993; Stryker 1991) and psychophysical (Wallis and Bülthoff 2001) studies have shown that these correlations influence the learning of object representations.

Several mechanisms have been proposed for how these correlations could be used for learning invariant representations (Becker 1993; Einhäuser et al. 2002; Földiák 1991; Stringer et al. 2006; Wallis 1996; Wiskott and Sejnowski 2002). Földiák (1991) proposed a modified Hebbian learning rule—the trace rule—that exploits temporal correlations in a sequence of input patterns. The trace learning rule has been used in a hierarchical multilayer network, to achieve invariant response properties for more realistic stimuli (Rolls and Stringer 2006; Stringer and Rolls 2002; Wallis and Rolls 1997).

How the trace rule is implemented in cortical circuits is still an open question. Wallis and Rolls (1997) argued that persistent firing, the binding period of glutamate in the N-methyl-D-aspartate (NMDA) channels, or postsynaptically released chemicals such as nitric oxide might be the biological basis for the trace rule. Sprekeler et al. (2007) showed theoretically that the learning rule for slow feature analysis (SFA), which is related to trace learning, can be achieved with spiking neurons. Nevertheless, invariance learning on the basis of temporal correlations has not yet been implemented in a network of spiking neurons.

Previous models for invariance learning (Einhäuser et al. 2002; Riesenhuber and Poggio 1999; Wallis and Rolls 1997) relied on not only the learning of features but also learning the specific connections to pool across related features to achieve invariant representations. We will show that feature representations can be learned in an ordered way, such that related features are represented in a local neighborhood and invariance can be achieved by a generic connectivity without the need for further learning. The key mechanism for this is to learn a topographic map that reflects the spatiotemporal correlations of the inputs.

Topographic maps and spatiotemporal stimulus correlations

Many cortical areas are topographically organized. In primary visual cortex (V1), neighboring neurons receive input from neighboring parts of the retinal image. Superimposed on the retinotopic organization is an orientation topography: neighboring populations of neurons respond to edges of similar orientation (Hubel and Wiesel 1974). In inferotemporal cortex,
topography for more complex features or even for characteristics of object views was found (Wang et al. 1996). This suggests that some higher-order features of the input are mapped continuously in a topographic fashion (for review see Tanaka 1996, 2003).

The model for the self-organization of cortical maps proposed by von der Malsburg (1973) relies on Hebbian learning in forward connections, short-range lateral excitation, and long-range lateral inhibition. A biologically realistic implementation of this learning principle is the RF-SLlSSOM (receptive field-spiking laterally interconnected synergetically self-organizing map; Choe and Miikkulainen 1998) model, which uses spiking model neurons. Trained with a stimulus set of oriented bars, these models can learn orientation maps similar to those found in primary visual cortex. In these studies, stimuli were presented in pseudorandom order to exclude the effects of temporal correlations. As our results show, temporal correlations can affect the emerging topography in this model architecture, if the lateral connections have a large time constant.

An attempt to extend the von der Malsburg model to account for temporal correlations has been considered by Wiemer and colleagues (Wiemer 2003; Wiemer et al. 2000). It is based on lateral propagation of activity, but has not been implemented in a biologically realistic network.

Goals and hypotheses

In this study we investigate a learning principle that combines the idea of spatial and temporal correlation-based invariance learning with self-organizing map formation. Hebbian learning suggests that the emerging topography of a self-organizing network with slow lateral connections is influenced not only by spatial but also by temporal correlations (Saam and Eckhorn 2000). In this study our main hypothesis is that temporal correlations in input sequences can shape the neighborhood relations in a learned topographic map. Furthermore, we hypothesize that a feature topography that reflects spatial and temporal correlations can support the view-invariant codings of object identity. We investigated these hypotheses with simulations of a biologically plausible network of spiking neurons. The slowness principle for learning invariant representations can be implemented in a biologically realistic spiking neural network by using NMDA-mediated short-range lateral connections and long-range lateral inhibition. This connectivity can cause a network dynamics with persistent activity that implements a memory trace. By manipulating the temporal correlations of the input we systematically investigated the effects of stimulus similarity and temporal proximity. View invariance is achieved by neurons of a downstream area that receive input from the topographic map via fixed, generic connections.

METHODS

Network architecture

The network consists of a forward pathway of three layers of spiking neurons. Layer E0 is the input layer, layer E1 represents the map formation layer, and the output layer E2 represents a cortical stage downstream of layer E1 (Fig. 1). Neurons in layers E0 (30 × 30 or 8 × 24 × 26 neurons, depending on the stimulus set), E1 (100 × 100), and E2 (10 × 10) are arranged in two-dimensional (2D) arrays. E0 neurons are activated by the stimulus patterns (see following text). E0 has α-amino-3-hydroxy-5-methyl-4-isoxazolopropionic acid (AMPA)-mediated excitatory forward projections (\(W_{E1,E0}\)) to the excitatory neurons of layer E1. These connections exhibit Hebbian plasticity. The connectivity from E0 to E1 is initially all-to-all with equal weights.

In addition to input from E0, E1 neurons receive excitatory input from their neighbors (\(W_{E1,E1}\)) with fixed connection strengths that decrease with the distance between two neurons according to a Gaussian

\[
w(E1, E1)_{ij} = \begin{cases} S_{E1,E1} \exp \left( -\frac{d_{ij}^2}{2\sigma_{E1,E1}^2} \right) & i \neq j \\ 0 & i = j \end{cases} \]

where \(w(E1, E1)_{ij}\) is the synaptic strength (weight) of the connection from neuron \(j\) to neuron \(i\), \(S_{E1,E1}\) is the maximum connection strength, \(d_{ij}\) is the Euclidean distance between neurons \(j\) and \(i\), and \(\sigma_{E1,E1}\) is the width of the Gaussian kernel. We used toroidal boundary conditions to avoid boundary effects. E1 neurons mutually inhibit each other via a pool of inhibitory interneurons I1. The connectivity between E1 and I1 is random; thus the pool of inhibitory neurons (I1) has no topographical order. Lateral excitatory connections from E1 to E1 and from E1 to I1 are mediated via fast AMPA (\(\tau_{decay} = 2.4\) ms) and slow NMDA (\(\tau_{decay} = 100\) ms) currents. Inhibitory connections from I1 to I1 and I1 to E1 are mediated by a γ-aminobutyric acid type A receptors.
(GABA<sub>A</sub>) current (fast, \(\tau_{\text{decay}} = 7.0\) ms). E1 neurons project to output layer E2 with a Gaussian weight profile

\[
w(E2, E1)_{ij} = S_{E2,E1} \exp \left( -\frac{1}{2} \left( \frac{d_{ij}}{\sigma_{E2,E1}} \right)^2 \right) \tag{2}\]

These connections were fixed and did not change during the simulation. Thus a neuron in layer E2 receives input from a fixed, localized region of layer E1. The connectivity patterns are summarized in Table 1.

**Model neurons**

Spiking neurons were simulated by a standard leaky integrate-and-fire model with a voltage threshold and biologically realistic synaptic potentials (Brunel and Wang 2001; Deco and Rolls 2005)

\[
C_m \frac{dV(t)}{dt} = g_L(V(t) - E_L) - I_{syn}(t) \tag{3}
\]

where \(C_m\) is the membrane capacitance and \(g_L\) is the leak conductance of the membrane. When the membrane potential exceeds the firing threshold \(\Theta\), an action potential (spike) is generated. The downstroke of the spike is modeled by resetting the membrane potential to \(V_{\text{reset}}\). After each spike an absolute refractory period of 1 ms duration is introduced. Parameter values are given in Table 2.

Excitatory forward connections are mediated by AMPA currents, lateral excitatory connections are mediated by AMPA and NMDA currents, and inhibition is mediated by fast GABA<sub>A</sub> currents. \(I_{syn}(t)\) is the sum of the AMPA, NMDA, and GABA<sub>A</sub> synaptic currents

\[
I_{syn}(t) = I_{\text{AMPA}}(t) + I_{\text{NMDA}}(t) + I_{\text{GABA}}(t) \tag{4}
\]

\[
I_{\text{AMPA}}(t) = g_{\text{AMPA}}(t)G_{\text{AMPA}}[V(t) - E_{\text{AMPA}}] \tag{5}
\]

\[
I_{\text{GABA}}(t) = g_{\text{GABA}}(t)G_{\text{GABA}}[V(t) - E_{\text{GABA}}] \tag{6}
\]

\[
I_{\text{NMDA}}(t) = G_{\text{NMDA}}[V(t) - E_{\text{NMDA}}]/\left(1 + [Mg^{2+}]\exp(-0.062V(t))/3.57\right) \tag{7}
\]

where \(E_{\text{AMPA}} = 0\) mV, \(E_{\text{NMDA}} = 0\) mV, and \(E_{\text{GABA}} = -70\) mV are the reverse potentials for the synaptic currents. The nonlinear voltage dependence of the NMDA current (caused by the Mg<sup>2+</sup>-block, Eq. 7) was modeled according to Jahr and Stevens (1990).

\(G_{\text{AMPA}}, G_{\text{GABA}},\) and \(G_{\text{NMDA}}\) are the maximum synaptic conductivities when all channels are open. \(g_{\text{AMPA}}(t), g_{\text{GABA}}(t),\) and \(g_{\text{NMDA}}(t)\) are the respective fractions of open channels. When a presynaptic spike occurs at \(t = t_{\text{sp}}\), the fraction of open channels \(G(t)\) increases and then decreases. This process is modeled with a difference of two exponentials (Eq. 8)

\[
G(t) = G_{\text{rise}}(t) - G_{\text{decay}}(t) \tag{8}
\]

\[
\frac{d}{dt}G_{\text{rise}}(t) = -\frac{G_{\text{rise}}(t)}{\tau_{\text{rise}}} + \frac{w_{nm}e_{nm}(t)\delta(t - t_{\text{sp}})}{\tau_{\text{sp}}} \tag{9}
\]

**Table 1.** Connection properties

<table>
<thead>
<tr>
<th>Connection</th>
<th>Connectivity Schema</th>
<th>Postsynaptic Currents</th>
</tr>
</thead>
<tbody>
<tr>
<td>E0 (\rightarrow) E1</td>
<td>All-to-all, modifiable</td>
<td>AMPA (fast)</td>
</tr>
<tr>
<td>E1 (\rightarrow) E1</td>
<td>Gaussian kernel with range (\sigma_{E1E2})</td>
<td>AMPA (fast) + NMDA (slow)</td>
</tr>
<tr>
<td>E1 (\rightarrow) I1</td>
<td>Random, connectivity (= c_{E1I1})</td>
<td>AMPA (fast) + NMDA (slow)</td>
</tr>
<tr>
<td>I1 (\rightarrow) E1</td>
<td>Random, connectivity (= c_{E1I1})</td>
<td>GABA&lt;sub&gt;A&lt;/sub&gt; (fast)</td>
</tr>
<tr>
<td>I1 (\rightarrow) I1</td>
<td>Random, connectivity (= c_{E1I1})</td>
<td>GABA&lt;sub&gt;A&lt;/sub&gt; (fast)</td>
</tr>
<tr>
<td>E1 (\rightarrow) E2</td>
<td>Gaussian kernel with range (\sigma_{E2E2})</td>
<td>AMPA (fast)</td>
</tr>
</tbody>
</table>

Connectivity and postsynaptic currents are shown for all synaptic connections between neuron layers. Connections between E0 and E1 are modifiable (see text), whereas all other connections are fixed.

**Table 2.** Model neuron parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Excitatory Neurons</th>
<th>Inhibitory Neurons</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_m)</td>
<td>0.5 nF</td>
<td>0.2 nF</td>
</tr>
<tr>
<td>(g_L)</td>
<td>25 nS</td>
<td>20 nS</td>
</tr>
<tr>
<td>(\Theta)</td>
<td>-50 mV</td>
<td>-50 mV</td>
</tr>
<tr>
<td>(V_{\text{reset}})</td>
<td>-55 mV</td>
<td>-55 mV</td>
</tr>
</tbody>
</table>

Parameters for inhibitory and excitatory neurons were taken from Deco and Rolls (2005).

\[
\frac{d}{dt}G_{\text{decay}}(t) = -\frac{G_{\text{decay}}(t)}{\tau_{\text{decay}}} + \frac{w_{nm}e_{nm}(t)\delta(t - t_{\text{sp}})}{\tau_{\text{sp}}} \tag{10}
\]

where \(w_{nm}\) is the synaptic weight and \(e_{nm}\) is the synaptic efficacy. The forward connections from E0 to E1 are not depressive \([e_{nm}(t) = \text{const} = 1]\) and evoke only an AMPA current. The recurrent connections between E1 neurons evoke both AMPA and NMDA currents. The ratio between the peak amplitude of NMDA and AMPA currents was set to 0.3 (Cairn and Malenka 1995). These recurrent connections show synaptic depression to stabilize the network activity. For the synaptic dynamics we used a simplified version of the model proposed by Tsodyks et al. (1998)

\[
\frac{d}{dt}e_{nm}(t) = \frac{1 - e_{nm}(t)}{\tau_{\text{rec}}} - U_{\text{in}}e_{nm}(t_{\text{sp}})\delta(t - t_{\text{sp}}) \tag{11}
\]

where \(U_{\text{in}}\) is the fraction of available transmitter that is released during a postsynaptic spike and \(\tau_{\text{rec}}\) is the recovery time constant for the transmitter pool.

**Learning rule**

We used a Hebbian learning rule similar to that proposed by Gerstner et al. (1996), Saam and Eckhorn (2000), and Michler et al. (2006). The synaptic weights \(w_{nm}\) of the forward connections from layer E0 to E1 are adapted according to the following equations

\[
\frac{d}{dt}w_{nm} = \delta_{nm}(t)RL_{\text{pre},m}L_{\text{post},m} \tag{12}
\]

\[
L_{\text{pre},m} = \sum_n \exp\left(-\frac{t - t_{\text{pre}}}{\tau_{\text{pre}}}\right) \tag{13}
\]

\[
L_{\text{post},m} = \sum_n \exp\left(-\frac{t - t_{\text{post}}}{\tau_{\text{post}}}\right) \tag{14}
\]

where \(\delta_{nm}(t)\) is 1 when a spike occurs in the postsynaptic neuron \(m\); otherwise, \(\delta_{nm}(t)\) is zero. \(t_{\text{pre}}\) and \(t_{\text{post}}\) denote the times of the past pre- and postsynaptic spikes. When a spike occurs, the pre- or postsynaptic learning potentials \(L_{\text{pre},m}\) or \(L_{\text{post},m}\) are increased by 1. They exponentially decrease with time constants \(\tau_{\text{pre}} = 20\) ms and \(\tau_{\text{post}} = 10\) ms. The exact values of these parameters are not critical. \(R\) corresponds to the learning rate. Because learning occurs only after postsynaptic spikes \([\delta_{nm}(t) = 1]\), this learning rule is temporally asymmetric: it prefers presynaptic before postsynaptic spiking. The learning rule increases weights if pre- and postsynaptic neurons have overlapping spike trains on a short timescale on the order of \(\tau_{\text{pre}}\) and \(\tau_{\text{post}}\).

Each time the firing rate of a postsynaptic neuron exceeds a threshold \((\theta_{\text{norm}} = 50\) Hz), all input weights are multiplied by normalization factor \(f_{\text{norm}} < 1\). Evidence for normalization of synaptic weights exists (e.g., Royer and Paré 2003), but the mechanisms are not yet understood. Weight normalization prevents infinite growth of weights and introduces competition between the inputs of a neuron.
Stimuli

The network was trained with sets of parameterized stimuli that differed along two parameter dimensions, denoted X and Y, respectively. We tested three increasingly complex stimulus sets, with different correlation structures.

GAUSSIAN STIMULI. Gaussian stimuli consisted of 2D Gaussian activity profiles varying in the horizontal and vertical positions of the center of the Gaussian. These coordinates were used as X and Y dimensions of the stimulus space. The correlation structure of this stimulus set is symmetrical in X and Y. Because of this symmetry we can isolate the effects of the temporal correlations by using stimulus sequences with temporal correlations along either the X or the Y direction of the stimulus space.

PRISM STIMULI. We generated a set of stimuli with variation corresponding to viewing angle (X parameter) and object identity (Y parameter) of three-dimensional (3D) objects. Objects were triangular prisms (Fig. 2A).

We varied an arbitrary set of parameters of the prism: the height, the size of the top and bottom triangles, the rotation angle between the top and the bottom triangles, and 3D orientation of the top and bottom triangles. Each of these parameters was systematically changed in steps according to a periodic triangular function \( \phi(\text{Y+}d) \) (Fig. 2B), which maps the parameter values to the Y dimension of the 2D stimulus parameter space. Therefore the shape changed only along a one-dimensional manifold. Shifting the phase of the triangular function \( \phi \) for different parameters, we obtained toroidal boundary conditions for the stimulus deformation parameter Y. An irregular texture was applied to the surfaces of the prisms to make the faces of the prism more distinct (Fig. 2C). Using the open source 3D library Crystal Space (Tybergehein et al. 2007) we generated views of these objects, rotated around their vertical axis with a step size of 18°, resulting in a set of 20 \( \times \) 20 stimulus pictures, each with 200 \( \times \) 200 pixels (Fig. 2D). Stimuli were preprocessed by a set of 30 \( \times \) 30 Gabor filters of 19-pixel wavelength and 6-pixel width of Gaussian, comprising eight orientations. To reduce the number of input neurons required, the resolution of the input array was reduced to 30 \( \times \) 30 by resampling and cropped to 26 \( \times \) 24 pixels. The outputs of these orientation filters were then used as input signals for the E0 neurons.

COIL STIMULI. To test the performance of the network for more natural stimuli we used images of natural objects taken under different viewing angles. Images were taken from the Columbia Object Image Library (COIL-100) database (Nene et al. 1996). We created a stimulus set with 10 objects and 36 views of each object. The X dimensions corresponded to the viewing angle and the Y dimension to object identity. With respect to the prism stimulus set, the pictures were preprocessed by 30 \( \times \) 30 Gabor filters (eight orientations; 10-pixel wavelength; 2.1-pixel width of Gaussian), resampled to 30 \( \times \) 30 pixels, and cropped to 26 \( \times \) 24 pixels. In contrast to the Gaussian and prism stimuli, in this stimulus set there was no continuous transformation along the Y dimension (object identity) of the stimulus space.

Training and test procedures

We used three training conditions with different temporal correlations between the elements of the stimulus set. In the X slow condition the X parameter was held temporally constant for intervals of \( t_{\text{const}} \) \( [400 \text{ ms}, 600 \text{ ms}] \), whereas the Y parameter was changed continuously. After each of these training intervals, a short interstimulus interval (20 ms) occurred and X and Y parameters switched to random values for the next training interval (see Supplemental Fig. S1).1 In the Y slow condition temporal correlations were conversed: the Y parameter was held constant for durations of \( t_{\text{const}} \) whereas the X parameter changed continuously. Thus temporal correlations were restricted to the fast changing dimension of the stimulus set.

As a control we simulated a random training condition with random order of stimuli in the sequence, i.e., without temporal correlations. Network simulations were performed with 125-s training epochs in alternation with test epochs. Both training and testing were done with the full stimulus sets. With 20 training epochs for the Gaussian and prism stimuli and 10 training epochs for the COIL stimuli, total simulated training times were 2,500 and 1,250 s, respectively. During the training epochs the forward connections from layer E0 to E1 were adapted according to the Hebbian learning rule.

During test epochs we tested the network properties with the complete stimulus set. Hebbian plasticity was turned off. Each stim-

---

1 The online version of this article contains supplemental data.

---

Fig. 2. Three-dimensional (3D) stimulus set. A: triangular prism. B: periodic triangular function used to continuously change the 3D object parameters along the Y-axis of the stimulus space. C: surface texture of the prism. D: a 3D-object stimulus set was generated by deforming and rotating the prism (see text).
Stimuli in this stimulus space is the shortest distance along a circular distance between stimuli 0 and 19 is 1. The difference between two dimensions of the stimulus space. The stimulus space is circular (e.g., distance between stimuli 0 and 19 is 1). The difference between two stimuli in this stimulus space is the shortest distance along a circular path

\[ e_x = \min (|X_{n} - X_{0}|, |X_{n} + N_s - X_{0}|, |X_{n} - N_s - X_{0}|) \]  

(15)

The estimated X value \( X_{\text{est}} \) is calculated by taking into account the activity of all E2 neurons (\( a_{j}[n], j \in \{1, \ldots, J_{E2}\} \)), elicited by the current stimulus, and the corresponding tuning curves \( T[X, j] \) of the E2 neurons. For a given value X the neural activity of a single neuron j multiplied by the corresponding value of the tuning curve \( T[X, j] \) is a measure for how strong this neuron estimates value X. The sum of this measure over all neurons is the population prediction \( P[X] \)

\[ P[X] = \sum_{j=1}^{N_s} (T[X, j]a_{j}[n]) \]

(16)

The estimated value \( X_{\text{est}} \) is the one with the highest likelihood

\[ X_{\text{est}} = \arg \max \{ P[X] \} \]

(17)

The tuning curve \( T[X, j] \) is calculated using the E2 responses of the penultimate test epoch \((n-1)\)

\[ T[X, j] = \frac{1}{N_s} \sum_{n=0}^{N_s-1} a_{n-1}[X, Y, j] \]

(18)

The original preference indices are in the range from 0 to 19. Because of the toroidal boundary conditions values 0 and 19 are direct neighbors in stimulus space. Therefore the maximal difference is 10. Note that for a uniform distribution, estimation error values of 0 and 10 would have a probability of 5%, whereas because of the rectification (Eq. 15), the values 1 to 9 would have a probability of 10%.

For a representation that is invariant with respect to the X parameter and selective for the Y parameter, the mean estimation error for the Y parameter \( e_y \) would be low and the mean estimation error for the X parameter \( e_x \) would be high. If the E2 neurons contained no information about the X parameter of the stimulus, the X estimation error would be uniformly distributed.

**RESULTS**

**Formation of topographic maps**

After training with the Gaussian stimulus set, all layer E1 neurons responded selectively to a small subset of the stimuli. Figure 3A shows the response matrix for a typical layer E1 neuron after training with the Gaussian stimulus set. The neuron encodes a continuous subregion of the stimulus space.

To quantify the selectivity, we calculated the mean response for each combination of X and Y stimulus parameter values. To visualize the spatial distribution of the stimulus selectivities, we represented the preferred X and Y parameters of each layer E1 neuron by the hue and the maximal response strength by the brightness of HSV (Hue, Saturation, Value) color space. Figure 4, A and B shows the topographic maps that were learned with the Gaussian stimuli, using the X dimension as the slow parameter and the Y dimension as the fast changing parameter. Both maps show patches of neighboring neurons with the same or similar selectivities. However, the patches are larger for the X parameter (Fig. 4A) than those for the Y parameter (Fig. 4B). Moreover, neurons with a preference for the same X parameter are clustered within a single local region of the map. In contrast, patches of neurons with a preference for a certain Y parameter value are distributed across the map.

These properties of the maps are exchanged when the temporal correlations of X and Y parameters are exchanged: Fig. 4, E and F shows the maps that were learned with the Y dimension as the slowly changing parameter. Here the patches of similar Y preference are larger and localized (Fig. 4F), whereas the representation of the X parameter (Fig. 4E) shows smaller patches and is more distributed across the map, showing a pattern similar to the pinwheel topography of V1 orientation selectivity (Bonhoeffer and Grinvald 1991). We see that in both cases similar values for the slow parameter (Fig. 4, A and F) are represented in a localized part of the map, whereas the fast changing parameter has a distributed representation (Fig. 4, B and E). In many cases the whole range of preferences for the fast changing parameter can be found within a patch of similar preference for the slow parameter.

**FIG. 3.** Learned stimulus selectivities. Stimulus response matrix and X and Y tuning curves for an example layer E1 neuron (A) and a layer E2 neuron (B). The maxima of the X and Y tuning curves are defined as preferred X and Y indices. The response matrices show the response strength for each of the 400 stimuli. The X and Y tuning curves are a measure for the selectivity to the 2 dimensions of the stimulus set. Here, Y was the slowly changing parameter. A: the E1 neuron encodes a subregion of the stimulus space. B: response of the E2 neuron showed high selectivity for the Y parameter and low selectivity for the X parameter.
In the condition of random presentation (Fig. 4, C and D), there were no qualitative differences between the maps for the preferred X and Y parameters. The topographic maps obtained with the prism and COIL stimuli (not shown) looked similar to those of the Gaussian set. To quantitatively compare the patch structure of the different maps, we calculated the Fourier spectra of the topographic maps and used the peak spatial frequency as an estimate of the patch sizes (Table 3). For all simulations with the Gaussian or prism stimuli and X as the slow parameter, the peak spatial frequency for the X parameter was much lower than that for the Y parameter. Conversely, when Y was the slow parameter the peak spatial frequency was lower in the Y map. We can conclude that the topographic maps for the slow parameter show larger patches compared with the maps for the fast changing parameter. This indicates that the temporal correlations are reflected in the learned topography. For the COIL stimuli, in the X slow condition the difference in patch sizes is very small. This is caused by the strong asymmetry in spatial correlations between X and Y dimensions of the stimulus space: strong correlations in the X direction (same object, different viewing angle), low correlations in the Y direction (same viewing angle, different object).

To illustrate the topographic order, we determined the regions in the map activated by the same object for different viewing angles (Fig. 5). A patch of high neural activity is continuously shifted as the viewing angle of the object changes, similar to activity in inferotemporal cortex evoked by different views of a face (Wang et al. 1996). Different views of the same object are mapped in the same region and have overlapping representations.

### Stability of learned preference maps

To investigate the convergence of the learned representations we performed an analysis of the temporal development of the learned preference maps in a simulation with 20 training epochs of 250 s and with Y as the slow parameter. For each neuron, we calculated the differences between X and Y preference values in each epoch to the respective preference values after the following training epoch. The fraction of neurons with a difference $>1$ decreased from 62% to 11% for the X preference and from 31% to 4.5% for the Y preference. Both

---

**TABLE 3. Spatial frequencies of topographic maps (Fig. 4), normalized to the dimensions of layer E1 (100 x 100)**

<table>
<thead>
<tr>
<th>Stimulus Set</th>
<th>X s.f.</th>
<th>Y s.f.</th>
<th>X s.f.</th>
<th>Y s.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>0.98</td>
<td>2.16</td>
<td>2.11</td>
<td>0.98</td>
</tr>
<tr>
<td>Prism</td>
<td>1.00</td>
<td>2.11</td>
<td>1.62</td>
<td>1.10</td>
</tr>
<tr>
<td>COIL</td>
<td>1.76</td>
<td>1.87</td>
<td>1.95</td>
<td>0.98</td>
</tr>
</tbody>
</table>

In all cases the dominant spatial frequency (s.f.) is lower for the slow and higher for the fast and continuously changing parameter. For the COIL stimulus in the X slow condition the differences are very small because of the biased correlation structure of this stimulus set (see text).
we calculated a selectivity index: the $Y$ parameter and more invariant for the $X$ parameter. Thus the response of this neuron is more selective for the $X$ tuning curve shows much larger variance than the $X$ tuning curve. There is a clear asymmetry. The $Y$ layer E1 neuron (Fig. 3B) was represented by a Gaussian stimuli. Compared with the response matrix of a layer E1 neuron (Fig. 3A) there is a clear asymmetry. The $Y$ tuning curve shows much larger variance than the $X$ tuning curve. Thus the response of this neuron is more selective for the $Y$ parameter and more invariant for the $X$ parameter.

From the minima and maxima of the $X$ and $Y$ tuning curves we calculated a selectivity index: $s = (\max - \min) / (\max + \min)$, in which $s$ measures the relative difference in responses to different stimulus patterns and is zero for a flat tuning curve. The $X$ and $Y$ selectivity index values for the layer E2 neurons are plotted against each other in the diagrams in Fig. 6. Figure 6A shows results for the Gaussian stimuli. In the simulation with $X$ as the slow parameter, $X$ selectivity of layer E2 neurons is higher than $Y$ selectivity (triangles). Thus the network response is more selective for the slow $X$ parameter and more invariant with respect to the continuously changing $Y$ parameter. The pattern is reversed for the simulation with $Y$ as the slow parameter (diamonds).

The results are very similar for the simulations with the prism stimulus set (Fig. 6B), despite different spatial correlations in the stimulus sets. For the COIL stimuli, results for the $Y$ (object identity) slow condition are similar (Fig. 6C). In the $X$ (viewing angle) slow condition the distribution selectivity indices are near the diagonal (similar $X$ and $Y$ selectivities), slightly shifted toward higher $Y$ selectivity. This reflects the strong asymmetry in spatial correlations in the COIL stimuli.

Invariant representations

Patches representing the slowly changing parameter were larger than the patches for the fast changing parameter (Fig. 4). Specifically, the localized region corresponding to the patch representing a given value of the slowly changing parameter contained patches of all values of the fast changing parameter. As a consequence of this topography, neurons in E2, each receiving input from a local region of the map layer, showed selectivity for the slow parameter and invariance to the fast changing parameter. Figure 3B shows the response matrix and the $X$ and $Y$ tuning curves for an example layer E2 neuron for the Gaussian stimuli. Compared with the response matrix of a layer E1 neuron (Fig. 3A) there is a clear asymmetry. The $Y$ tuning curve shows much larger variance than the $X$ tuning curve. Thus the response of this neuron is more selective for the $Y$ parameter and more invariant for the $X$ parameter.

From the minima and maxima of the $X$ and $Y$ tuning curves we calculated a selectivity index: $s = (\max - \min) / (\max + \min)$, in which $s$ measures the relative difference in responses to different stimulus patterns and is zero for a flat tuning curve. The $X$ and $Y$ selectivity index values for the layer E2 neurons are plotted against each other in the diagrams in Fig. 6. Figure 6A shows results for the Gaussian stimuli. In the simulation with $X$ as the slow parameter, $X$ selectivity of layer E2 neurons is higher than $Y$ selectivity (triangles). Thus the network response is more selective for the slow $X$ parameter and more invariant with respect to the continuously changing $Y$ parameter. The pattern is reversed for the simulation with $Y$ as the slow parameter (diamonds).

When we used viewing angle ($X$) as the slow parameter the picture is reversed (Fig. 7C): $X$ estimation errors were low and $Y$ estimation errors were nearly uniformly distributed. Thus in this learning condition the network has learned a representation that can effectively code for the viewing angle but is invariant with respect to object identity. Note that the $X$ error distribution has a second peak at error value 7 (visible in Fig. 7, B and C), which is caused by the rotation symmetry of the prism stimulus.

For comparison we repeated the simulations with a random order of stimulus presentation. Thus there were only spatial and no temporal correlations. Figure 7B shows the estimation errors for this learning situation. The peaks in the distributions
for low errors are much smaller and reflect the spatial correlations in the stimuli.

Parameter variations

To test the robustness of the learning mechanism, we systematically varied stimulus timing and the properties of excitatory and inhibitory lateral connections. For these tests we used the Gaussian stimuli with Y as the slow parameter. To evaluate the network performance we defined test trials with an estimation error \( e_{Y} \) as correct predictions and the proportion of correct predictions as the performance. These performance values were plotted against the variations of simulation parameters in Fig. 8. Strong invariance is indicated by a high value in the Y performance and low value in the X performance because Y was the slow parameter. Chance level is 3/20 = 0.15.

We varied range and strength of lateral excitatory connections \( a_{E1,E1}, b_{E1,E1} \), strength of lateral inhibition \( b_{E2,E1} \), and the stimulus timing \( t_{stim} \). Figure 8, A–C shows the dependence of the X and Y performance on the range of the lateral excitatory connections for three different stimulus timing conditions \( t_{stim} \in \{10, 20, 40 \text{ ms}\} \). The network shows high invariance in a range \( 4 \leq a_{E1,E1} \leq 5 \) for all three stimulus timing conditions (Fig. 8, A–C). In Fig. 8D the stimulus timing was varied in the range \( 5 \text{ ms} \leq t_{stim} \leq 100 \text{ ms} \), whereas all other parameters were constant. For long stimulus presentation times of \( t_{stim} > 70 \text{ ms} \) the performance for the fast and the slow parameters were very similar around 0.5, and thus the responses in layer E2 showed no invariance.

When the strength of the lateral connections between E1 neurons was varied, the network showed high performance in a range \( 0.07 \leq b_{E1,E1} \leq 0.13 \) (Fig. 8E). Without the lateral connections \( b_{E1,E1} = 0 \), performance dropped to chance level. For weights >0.15 performance decreased as well. Thus although these ranges were fairly broad, lateral extent and strength of the lateral excitation should be within a proper range, corresponding to relative changes by a factor of 2. In contrast, the strength of the lateral inhibitory connections is uncritical (Fig. 8F). Network performance is very robust against increased inhibition over a wide range. Likewise, varying the time constants of the learning rule, \( \tau_{pre} \) and \( \tau_{post} \) by a factor of 2 from 10 to 20 ms, did not lead to qualitatively different results (data not shown).

The emergence of topographic maps in our model critically depends on persistent activity in localized groups of neurons, which acts as a memory trace. Figure 9A shows how the size of the activity patches representing the stimuli depends on the parameters of the lateral connectivity. Patch size increases with

---

**FIG. 7.** Estimation errors of layer E2 activity for simulation with the prism stimuli. A: Y (object identity) was the slow parameter, X (viewing angle) the fast changing parameter. B: random order of stimuli. C: X was the slow and Y the fast changing parameter.

**FIG. 8.** Effects of model and stimulus parameters on network performance. The network was trained with the Gaussian stimuli with Y as slowly changing parameter. The diagrams show the dependence of the X and Y performance on the range of the lateral connections in E1 \( b_{E1,E1} \) and stimulus sequence speed \( t_{stim} \). An invariant representation is indicated by high performance and low X performance. With a lateral interaction range \( \sigma = 4 \) the network learned invariant representations for a wide range of stimulus speed values. A: fast \( t_{stim} = 10 \text{ ms} \). B: \( t_{stim} = 20 \text{ ms} \). C: \( t_{stim} = 40 \text{ ms} \). D: with increasing stimulus duration \( t_{stim} \), Y performance increases and X performance drops. E: strength of the lateral connections in E1 \( b_{E1,E1} \) was varied. F: strength of the lateral inhibitory connections \( b_{I1,E1} \) was varied.
larger lateral excitation range $\sigma_{EI,E1}$ and decreases with stronger lateral inhibition. However, the strength of lateral excitatory connections, determined by the amplitude of the Gaussian kernel $S_{EI,E1}$, did not influence the size of activity patches. Patch size in turn influenced the learned stimulus preference maps. As Fig. 9B shows, larger patch size leads to maps with lower spatial frequency.

**Discussion**

We investigated a mechanism for learning invariant properties of input stimuli. This mechanism implements the idea of extracting slowly varying features from input sequences. It can be applied for learning invariant representations of visual objects. When view-variant retinal projections of an object are presented successively, the spatiotemporal correlations in the input lead to a locally connected, restricted representation in a topographic map. This topographic representation can be used to produce invariant responses in neurons at a successive stage, without further learning, via a simple, unspecific connection scheme. Our approach combines the principles of invariance learning by exploiting temporal correlations and self-organization of topographic maps. Furthermore, it demonstrates that learning of slowly varying features can be achieved in a network of spiking neurons, which is a necessary requirement for a biologically realistic mechanism. Furthermore, our results suggest a functional relevance of cortical topographic maps.

**Spatiotemporal input correlations and topographic maps**

The architecture of our network is similar to that proposed by von der Malsburg (1973). This architecture is an application of the principle of pattern formation by local self-enhancement and long-range inhibition (Gierer and Meinhardt 1972). The basic building blocks are adaptive, Hebbian forward connections, long-range lateral inhibition, and short-range lateral excitatory connections. Trained with a set of stimuli, such networks transform the spatial correlations between stimuli into spatial proximity of their representations in the emerging map (Choe and Miikkulainen 1997; Kohonen 1982; von der Malsburg 1973).

It is possible to learn view-invariant representations by using spatial correlations only (Stringer et al. 2006), but this requires that spatial correlations between different views of the same object are higher than spatial correlations between views of different objects. This is the case for our simulations with the COIL stimulus set. Even without temporal correlations along the object dimension, the strong spatial correlations along the viewing angle dimension and weak spatial correlations along the object dimension lead to selectivity for object identity. However, in many real-life viewing situations views of different objects (such as faces) can be highly correlated if seen from the same viewing angle, whereas different views of the same object can result in highly different retinal images. With such a stimulus set Wiemer (2003) observed emergence of selectivity for viewing angle. As with our COIL stimulus set, the spatial correlations in the stimulus set dominated the selectivity after learning.

Our prism stimulus set has correlations along both dimensions of the stimulus set (viewing angle and object identity). Under these conditions, spatial correlations alone are not sufficient to learn view-invariant representations that are selective for object identity. Therefore both spatial and temporal correlations must be exploited.

Under natural viewing conditions different views of the same visual object often occur in temporal proximity. We mimicked such viewing conditions by creating stimulus sequences with temporal correlations along only one dimension of the stimulus space. Many different models have been proposed for how these temporal correlations can be used for learning invariant representations of visual objects (Becker 1993; Einhäuser et al. 2005; Földiák 1991; Rolls and Stringer 2006; Stringer and Rolls 2002; Wallis and Rolls 1997; Wiemer 2003; Wiemer et al. 2000; Wiskott and Sejnowski 2002). Our study shows how a biologically plausible network of spiking neurons can make use of temporal correlations to achieve invariant representations.

In contrast to most models of self-organizing maps (e.g., Choe and Miikkulainen 1997; Erwin et al. 1995; Goodhill and Cimponeriu 2000; Goodhill and Willshaw 1990; Kohonen 1982; Swindale 1996; von der Malsburg 1973) in our simulations the network response to a stimulus depends not only on the learned forward connections, but also on the past activity of the map layer. A related principle has been investigated by Wiemer (2003). However, in this study, the relevance of the learned topography for invariant representations was not considered.

**Network dynamics and influence of parameters**

In previous models for invariance learning from temporal correlations (Einhäuser et al. 2005; Földiák 1991; Rolls and Milward 2000; Wiskott and Sejnowski 2002), the slowness principle was built into the learning rule. In our network, the synaptic learning rule operates only on a fast timescale. It cannot capture temporal correlations on a timescale much longer than 20 ms. Temporal input correlations on a longer
timescale are extracted by the network dynamics. Therefore the exact implementation of the learning rule—in particular, the pre- and postsynaptic terms—is uncritical. As Almassy et al. (1998) pointed out, a continuous firing of a local group of neurons has an effect that is similar to Földiák’s postsynaptic memory trace. In our network, persistent firing of local groups of E1 neurons is enabled by excitatory lateral interactions in layer E1, which are mediated by fast decaying AMPA currents and slowly decaying NMDA ($\tau_{\text{decay}} = 100 \text{ ms}$, Table 4) currents. These connections provide a local positive feedback, whereas the long-range inhibition reduces the activity in other parts of the layer. This is a neural implementation of the mechanism of biological pattern formation proposed by Gigerer and Meinhardt (1972). Note that for this mechanism to work in our case, the time constant of the slow excitatory component (NMDA) must be slower than the time constant of lateral inhibition (GABA). Otherwise, the lateral inhibition would synchronize the whole network and destroy competition between different parts of the map.

The combination of short-range lateral excitatory connections and long-range inhibition enhances activity differences in the E1 layer and results in a competitive network dynamics and local patches of activity can form. Furthermore, in the absence of E0 input, an activated local patch of neurons can keep its activity. This persistent activity is weakened by the depression mechanism in the excitatory lateral synapses. As a result, the patch of activity can move continuously in the E1 layer. Therefore stimuli that occur in temporal sequence—typically different views of the same object—tend to be represented in neighboring regions of the map (Fig. 5).

The specific network dynamics is an essential feature underlying the formation of the topography that captures spatiotemporal correlations. Thus according to our model, one would expect to find persistent activity of local groups of neurons in cortical areas with topographic maps. Furthermore, one would expect that features with similar spatial correlations are represented closer to each other if they are also temporally correlated. This could be tested in experiments investigating the selectivity to object stimuli (e.g., Logothetis et al. 1995) by varying the temporal correlations of the stimuli.

The size of an activity patch in the E1 layer mainly depends on the interaction of positive feedback from the activity center and negative feedback from global inhibition. It increases with longer lateral connections ($c_{\text{E1,E1}}$) and decreases by stronger lateral inhibition ($c_{\text{E1,E1}}$) (Fig. 9). Despite the dependence of network dynamics on several network parameters, our network is robust against changes in a wide range of parameters (Figs. 8 and 9).

In this study we considered the learning of topographic maps. Other parameters like the connectivity from E1 to E2 were fixed. To achieve invariant responses in layer E2 the convergence from layer E1 to E2 ($c_{\text{E2,E1}}$) must be in the range of the patch size in the topographic map for the slow stimulus parameter. Furthermore, we assume that network dynamics and learning rate are appropriate with respect to the typical time constants of changes in the inputs. In a biological network the relevant parameters would have to be adjusted by learning or evolutionary adaptation.

**Models of invariant representations**

An early approach for invariant object recognition is the dynamic routing model (Olshausen et al. 1993). In this model the visual input is transformed into a canonical, object-based reference frame. Although this mechanism can solve the problem of scale and translation invariance, it is insufficient for achieving view invariance because there is no simple geometric transformation between the front view and the back view of an object. Riesenhuber and Poggio (1999) proposed a hierarchical model that relies on the two alternating operations, template matching and pooling units (complex cells), and thereby achieve invariance over the corresponding subset of basic features. They suggest that the proposed connectivity could be learned with the trace rule (Földiák 1991). The VisNet model by Stringer and Rolls (2002) demonstrates how complex-cell connectivity can be learned from temporal correlations in continuous image sequences. Our model extends these approaches and further suggests a possible role of topographic maps for invariant object representations.

**Topographic representation and invariant responses**

As our results show, a topographic representation can be used to generate invariant responses by simple neural mechanisms. The invariance properties of the output layer (E2) neurons in our model (Fig. 3B) are a consequence of the topography in the map layer (E1) because E2 neurons receive input from a localized region in E1 and therefore represent the average activity in this region. After training with sequences of object views, neurons selective for different views of the same object are clustered in a local neighborhood in E1. Neurons in E2 average over such a neighborhood and thus their responses are invariant to viewing angle while maintaining selectivity for object identity. Thus invariance arises from the learned topography through a generic connection scheme without the need for further learning. Without a topography, to achieve invariance in E2 neurons would require specific connections from E1. Learning such specific connections is more costly because a higher number of initial connections must be provided. To achieve an invariant object representation from a population of feature coding cells, those cells must be selected that code for the same object. If these cells were randomly distributed (salt-and-pepper arrangement) in the previous processing layer, a high connectivity would be needed initially to ensure that there is at least one cell in the invariance layer that receives connections from all of them. Furthermore, another learning step would be required to achieve the adequate connectivity for invariant responses. In contrast, in our model invariant responses arise from averaging over a local neighborhood of the topographic map via fixed forward connections that need no further modifications.

The formation of cortical maps has been suggested to be the result of the minimization of wiring length between neurons processing related stimuli (Koulakov and Chklovskii 2001). Our approach is entirely compatible with this view because, in

### Table 4. Time constants for synaptic currents

<table>
<thead>
<tr>
<th>Synaptic Current</th>
<th>$\tau_{\text{rise}}$</th>
<th>$\tau_{\text{decay}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMPA</td>
<td>0.5 ms</td>
<td>2.4 ms</td>
</tr>
<tr>
<td>GABA_A</td>
<td>1.0 ms</td>
<td>7.0 ms</td>
</tr>
<tr>
<td>NMDA</td>
<td>5.5 ms</td>
<td>100.0 ms</td>
</tr>
</tbody>
</table>

\[J\text{ Neurophysiol} \cdot \text{VOL 102} \cdot \text{AUGUST 2009} \cdot \text{www.jn.org}\]
our simulations, the topographic maps emerge as a consequence of the assumption that lateral connections have limited length. In addition, our results demonstrate that the clustering of neurons with similar properties in these maps has the functional benefit that invariance with respect to certain stimulus dimensions can be achieved in a straightforward way.

Conclusions

We propose a mechanism for spatiotemporal correlation-based invariance learning that is compatible with the functional architecture and plasticity mechanisms in the cortex. Our network transforms spatiotemporal correlations of the input sequence into the topography of a self-organizing map. The activity in our network shows similarities to neural activity in inferotemporal cortex (IT), which contains a topographic representation of object features (Tanaka 1996, 2003). The basic mechanisms of our model exist in the ventral pathway of the visual cortex. Therefore it is feasible that the emergence of object feature topography in IT may be based on the principles proposed in our model.

The aim of this work, however, was not to model a specific cortical area. The invariance learning mechanism we described here could be at work for features of any complexity, at any stage in the cortical hierarchy, and in any sensory modality, corresponding to the widely observed occurrence of topographic maps in the cortex.

ACKNOWLEDGMENTS

We thank K. Tanaka for discussions, L. Wiskott and W. Einhäuser for valuable suggestions and critical remarks on an earlier version of the manuscript, and the anonymous reviewers for helpful comments. The Information Technology Department of the University of Marburg provided access to its modular architecture for robust computation cluster for running simulations.

GRANTS

This work was supported by Deutsche Forschungsgemeinschaft (DFG) Forschergruppe 560 and DFG Graduiertenkolleg 885.

REFERENCES


