LETTER **Communicated by Heinrich Buelthoff**

A Temporal Stability Approach to Position and Attention-Shift-Invariant Recognition

Muhua Li
limh@cim.mcgill.ca
James J. Clark
clark@cim.mcgill.ca
Centre for Intelligent Machines, McGill University, Montréal, Québec, Canada H3A 2A7

Incorporation of visual-related self-action signals can help neural networks learn invariance. We describe a method that can produce a network with invariance to changes in visual input caused by eye movements and covert attention shifts. Training of the network is controlled by signals associated with eye movements and covert attention shifting. A temporal perceptual stability constraint is used to drive the output of the network toward remaining constant across temporal sequences of saccadic motions and covert attention shifts. We use a four-layer neural network model to perform the position-invariant extraction of local features and temporal integration of invariant presentations of local features in a bottom-up structure. We present results on both simulated data and real images to demonstrate that our network can acquire both position and attention shift invariance.

1 Introduction

Humans are adept at visually recognizing objects or patterns under different viewing conditions. They are tolerant of position shifts, rotations, and deformations in the visual images. Psychological evidence (Bridgeman, Von der Hejiden, & Velichkovsky, 1994; Deubel, Bridgeman, & Schneider, 1998; Leopold & Logothetis, 1998; Norman, 2002; Walsh & Kulikowski, 1998) shows that there exist mechanisms along the visual pathway that maintain perceptual stability in the face of these variations in visual input. Hubel and Wiesel (1962) found that simple neurons in the primary visual cortex respond selectively to stimuli with specific orientation, while complex neurons present certain position-invariant properties. Neurons in higher visual areas, such as the inferotemporal cortex (IT), have larger receptive fields and show more complex forms of invariance. They respond consistently to scaled and shifted versions of the preferred stimuli (Gross & Mishkin, 1977; Ito, Tamura, Fujita, & Tanaka, 1995; Perrett, Rolls, & Caan, 1982; Rolls, 2000).

Maintaining perceptual stability is also an emerging issue in computer vision systems. An important consideration in the design of robotic vision systems is to be able to recognize the external world from the video stream acquired as the robot is wandering about. The video input is often erratic and unstable because the robot moves its eyes, head, and body to perceive the surroundings and avoid obstacles when it performs tasks. To perform well in recognition tasks, a robot should be able to maintain a constant perception of the structure of an object when changing views of the object during its motor activities.

A number of models have been proposed to describe the mechanisms underlying perceptual stability, such as spatial-phase invariance, translation invariance, and scale invariance (Chance, Nelson, & Abbott, 2000; Fukushima, 1980; Riesenhuber & Poggio, 1999; Salinas & Sejnowski, 2001). In particular, temporal association is deemed an important factor in the development of transformation invariance (Miyashita, 1988; Rolls, 1995). Temporal continuity was first employed by Földiák (1991) to capture the temporal relationship of input patterns. It has been demonstrated that transformation invariances such as translation or position invariance and viewpoint invariance can be learned by imposing temporal continuity on the response of a network to temporal sequences of patterns (Bartlett & Sejnowski, 1998; Becker, 1993, 1999; Einhäuser, Kayser, König, & Körding, 2002; Földiák, 1991; Körding & König, 2001).

The human visual system as a whole seamlessly combines retinal images and visual-related motor commands to give a complete representation of the observed external environment. However, most research work done so far has focused on achieving different degrees of invariance based only on the sensory input, while ignoring the important role of visual-related motor signals. In our opinion, visual-related self-action signals are crucial in learning spatial invariance, as they provide information as to the nature of changes in the visual input.

A critical issue that must be considered in modeling human vision is that the visual system has to deal with an overwhelming amount of information. It is well known that selective attention plays an important role in the human visual system by permitting the focusing on a small fraction of the total input visual information (Koch & Ullman, 1985; Maunsell & Cook, 2002). Shifting of attention enables the visual system to actively, and efficiently, acquire useful information from the external environment for further processing.

Our goal is to develop object recognition systems that use covert and overt shifts in attention for feature selection. Covert attention shifts result from a change in feature selection processes occurring with the eye held fixed. Overt attention shifts refer to the change in the image data being attended to that results from a large, saccadic eye movement. Both overt and covert attention shifts cause changes to the visual input that the object recognition system works on. It is important that the functioning of the object recognition system be invariant to the effects of these attention shifts.

With respect to changes induced by eye movements, or overt attention shifts, this invariance is specifically position invariance, where the recognition process should provide the same answer regardless of the location on the retina that the image of the object is projected.

Most object recognition techniques that employ attention shifts are mainly based on covert or overt attention shifts, and they rarely consider both. The bulk of these methods consider only covert shifts, where the retinal input to the systems remains unchanged during the learning and recognition process (Kikuchi & Fukushima, 2001; Olshausen, Anderson, & Van Essen, 1993). If we directly apply such methods to overt attention shifts, the distortions due to the nonuniformity of the retina and the nonlinearity of projection on to the hemispherical retina may cause problems when foveating eye movements take place. For example, even though Kikuchi and Fukushima's model of invariant pattern recognition (2001) employs a "scan path" of eye saccades, it does not model any associated relative distortions. In their approach, the only effect of eye movement is a spatial displacement of the imaged features. Their model achieves shift invariance and scale invariance based on extracted spatial relations, which are internally encoded by the visual system as a chain of saccadic vectors and fixated local features. This model is too simplistic, however; a true model of recognition with eye movements must take into account the image distortions resulting from eye movements.

In this letter, we propose a new approach to attaining position invariance in which the processes of covert and overt attention shifts play a central role. We implicitly assume that the variation of feature positions on various cortical feature maps arises entirely from the action of covert and overt attention shifts. Motion of scene features in the external world is irrelevant, as it is the action of the attention systems that determines the location of the scene features in the internal representations. In this way of thinking, position invariance is really invariance to attention shifts, whether they be covert or overt. Desimone (1990) points out that the effects on the visual cortex of covert and overt attention shifts are very similar. It is conceivable, therefore, that we could develop a unified approach in which covert and overt shifts are not distinguished. We employ a temporal difference learning scheme where knowledge of the attention shift command is used to gate the learning process, permitting temporal correlation to take place between visual inputs across attention shifts. We implement a four-layer neural network model and test it on simulated data consisting of various geometrical shapes undergoing transformations.

2 A Neural Network Model of Attention Shift Invariance

The overall model being proposed is composed of two submodules, as illustrated in Figure 1. One is the attention control module, which generates attention-shift signals according to a saliency map. This module also gen-

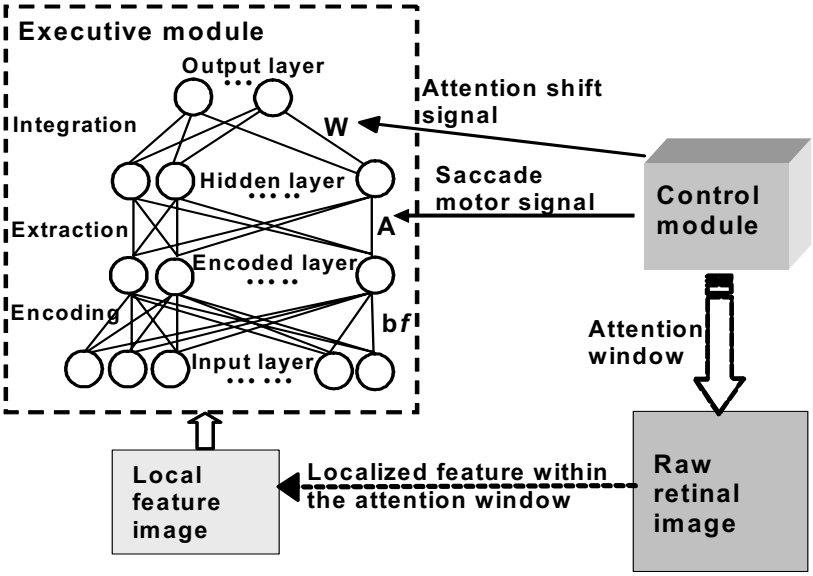


Figure 1: The proposed neural network model is composed of two modules: a control module and an executive module. The control module is an attention-shift mechanism that generates attention-shift signals and saccade motor signals to trigger the learning processes in the executive module. It also selects local features, which are part of the raw retinal image falling within the attention window, as input to the executive module. The executive module consists of a four-layer network, which accomplishes the extraction of position-invariant local features and the integration of attention-shift-invariant complex features from lower level to higher level, respectively.

erates saccadic motor command signals (or overt attention shift signals), which are used to determine the timing for learning. The module obtains as input local feature images from the raw retinal images via a dynamically position-changing attention window. The second submodule is the executive module, which performs the learning of invariant neural representations across attention shifts in temporal sequences. Two forms of learning, position-invariant extraction of local features, and integration of position-invariant object representation (a composition of a set of local features) across attention shifts, are triggered by the saccadic motor signals and attention shift signals from the control module, respectively.

2.1 Temporal Continuity Approaches to Development of Position Invariance. In this section, we detail how our network learns invariance to position changes that result from eye movements. Our approach is based on

the work of Földiák (1991) and Einhäuser et al. (2002). They proposed methods for developing position invariance that rely on a temporal continuity of the images of objects projected onto the retina. These methods have been shown to work well in learning position invariance in both simulated data and real-world video sequences. Einhäuser et al. proposed a three-layer feedforward network model capable of learning from natural stimuli that develops receptive field properties matching those of cortical simple and complex neurons. Hebbian learning in each output layer cell emphasizes the temporal structure of the input in the learning process. Experimental results show that the middle layer cells learn simple cell response properties that have strong selectivity to both orientation and position. The output layer cells learn complex cell response properties, which exhibit a level of position invariance while preserving orientation selectivity. However, their learning algorithm depends crucially on temporal smoothness in the input. The learning result is very sensitive to the timescale and the temporal structure in the input. When the time interval between successive input scenes becomes large, the temporal difference in the input data can also become large. Equivalently, when structures in the scene are moving rapidly, the temporal changes in the input stream can be large. Einhäuser et al. reported that the output layer cells lose the position-invariant property when the input lacks temporal smoothness.

In order to produce position-invariant recognition, the visual system must be presented with images of an object at different locations on the retina. In techniques such as those of Einhäuser et al. (2002) and Földiák (1991), it is mainly the motion of the objects in the external world that produces the required presentations of the object image across the retina. There are a number of problems with this. Most important, the motion of objects in 3D space can change the appearance of objects significantly. Thus, the problem of developing position invariance is converted to the much more difficult problem of developing viewpoint invariance. The difference in the appearance of a moving object is generally greater as the displacement increases. This means that only local position invariance can be learned.

In this letter, we propose a position-invariance learning method that is not overly affected by external object motion. The key aspect of our approach is the use of rapid attention shifts (overt or covert) to provide the necessary object image displacements. In this way, position invariance can be seen to arise from attention shift invariance. The short time between acquisition of images across an attention shift minimizes the change in the image due to motion of the object in space. Thus, in our approach, most of the change in the image is due to the attention shifts (and not to the motion of the object). Clearly, learning invariance with respect to some quantity requires exposure to data that varies only with respect to this quantity. Since we are learning invariance to attention shifts (covert or overt), we require a signal in which the variation is entirely due to attention shifts. This is accomplished by using only those images associated with attention shifts. At other times, the

images are changing due to extraneous factors, such as object motion, and thus can interfere only with the learning of the invariance. Our approach has the additional advantage that the attention shift command can be used as a signal to direct learning. We can, for example, arrange for learning to take place only during the period of time immediately before and after the attention shifts.

2.2 Extraction of Position-Invariant Local Features. The need for development of position invariance arises due to projective distortions and the nonuniform distribution of visual sensors on the retinal surface. These factors result in qualitatively different signals when object features are projected onto different positions of the retina. For example, when a linear object feature in space is projected onto a hemispherical surface, such as the retina, it is relatively undistorted when projected near the optical axis (i.e., near the fovea), whereas its image becomes curved when projected away from the optical axis (e.g., in the periphery). The problem of finding a position-invariant representation for such features can therefore be thought of as that of finding the underlying relationship between various distorted retinal images of the same physical feature at different retinal positions. We propose that this problem can be simplified if a canonical representation of the feature can be specified. The foveating capability of the human visual system gives us such a canonical representation. It is the role of the foveating system to shift the image of a feature being attended to in the retinal periphery to become centered on the fovea. The foveal image of an object feature is a suitable candidate for the feature's canonical representation since, statistically, among all the retinal images of a feature, the foveal image is the most frequently observed. Furthermore, the process of fixation and tracking ensures that the foveal image representation is very stable relative to the peripheral images. When we refer to the neural representation of a feature's foveal image as its canonical representation, the problem of position-invariant representation of a feature can be interpreted as one of associating the neural representations of all of its peripheral images with its single canonical representation.

At a deeper level, the approach that we are proposing involves executing self-actions of the observer (in this case, saccadic eye movements) and observing the resulting changes in the retinal image. The idea that knowledge of self-action and the resulting sensory changes plays a role in perception is becoming popular. For example, O'Regan and Noë (2001) proposed that visual percepts are based on the sensorimotor contingencies that describe the relation between motor activities and visual sensory input.

Our approach to the learning of position invariance is based on the proposal of Clark and O'Regan (2000) that position invariance could be achieved through learning of the sensorimotor contingencies associated with a given feature. They presented a prototype of an association mechanism using the temporal difference learning schema of Sutton and Barto

(1981) to learn the association between pre- and postmotor visual input data, leading to the desired position-invariance properties. A saccade is employed to foveate preattended features, so that associations between presaccadic peripheral stimuli and postsaccadic foveal stimuli (the canonical image) can be learned each time a saccade occurs. Given an input presaccadic neural response X, an association matrix V, and a reinforcement reward λ, their learning rule is as follows:

$$\Delta V_{ij} = \alpha(\lambda(t) + V_{ij}(t-1)[\gamma X_j(t) - X_j(t-1)])\bar{X}_j(t) \tag{2.1}$$

with

$$\Delta \bar{X}_j(t) = \delta(X_j(t-1) - \bar{X}_j(t-1)). \tag{2.2}$$

The reinforcement reward $\lambda(t)$ here is the postsaccadic neural response to the foveal feature (the canonical image), and has the same dimension as X.

Clark and O'Regan's model (2000) works well in handling geometric distortions of images features due to position variance. However, a limitation of their model is that the association is very space and time-consuming, with resource requirements growing exponentially with the number of input neurons.

In this letter, we provide a more efficient version of the Clark-O'Regan approach. Our aim is to reduce the computational requirements of their model while retaining the capability of learning position invariance of local features. We make a modification to their learning rule, using temporal differences over longer timescales rather than just over pairs of successive time steps. In addition, we use a sparse coding approach to reencode the simple neural responses, which reduces the size of the association weight matrix and therefore the computational complexity. Our model includes an input layer, an encoded layer, and a hidden layer, as well as the connection matrix between the layers.

We model the input layer neuronal receptive field profile with a Gabor-like function. We refer to the input layer units as simple neurons, as they have similar properties to the simple cells of the visual cortex. The response of each simple neuron to a retinal image is the convolution of its receptive field profile and the image. The simple neural responses then are encoded by a sparse coding approach (Hyvärinen & Hoyer, 2001; Olshausen & Field, 1996, 1997) to reduce the statistical redundancies in the input pattern. The learning of basis functions sets and their sparsely distributed coefficients ensures that only a small number of active neurons in the encoded layer represent the original input pattern. The details of the encoding process are as follows.

Let $\boldsymbol{F}$ denote the simple neuronal responses. A set of basis functions $\boldsymbol{bf}$ and a set of corresponding sparsely distributed coefficients $\boldsymbol{a}_i$ are learned

to represent $\boldsymbol{F}$:

$$F(j) = \sum_i a_i * bf_i(j) \Rightarrow F = bf * a. \tag{2.3}$$

The basis function learning process is a solution to a regularization problem that finds the minimum of a functional E. This functional measures the difference between the original neural responses $\boldsymbol{F}$ and the reconstructed responses $\boldsymbol{F}' = bf * a$, subject to a constraint of sparse distribution on the coefficients $\boldsymbol{\lambda}$:

$$E(bf, a) = \frac{1}{2} \sum_j \left[F_j - \sum_i a_i * bf_{ij} \right]^2 + \alpha \sum_i Sparse(a_i) \tag{2.4}$$

$$\text{where } Sparse(a) = \ln(1 + a^2). \tag{2.5}$$

Sparseness is enforced by the second term of equation 2.4, which drives the coefficients $\boldsymbol{a}$ toward small values.

In our implementation, $\boldsymbol{E}$ is minimized over its two arguments $\boldsymbol{bf}$ and $\boldsymbol{a}$, respectively. The minimization is first performed over $\boldsymbol{a}$, with a fixed value of $\boldsymbol{bf}$, and then performed over $\boldsymbol{bf}$.

The inner minimization loop over $\boldsymbol{a}$ is performed by iterating the nonlinear conjugate gradient method (Shewchuk, 1994) until the derivative of $\boldsymbol{E}(\boldsymbol{bf}, \boldsymbol{a})$ with respect to $\boldsymbol{a}$ is zero:

$$\frac{\partial E(bf_{ij}, a_i)}{\partial R_i} = \sum_j bf_{ij} * \left(F_j - \sum_k a_k * bf_{kij} \right) - \alpha * \frac{\partial Sparse(a_i)}{\partial a_i}. \tag{2.6}$$

The outer minimization loop over $\boldsymbol{bf}$ is accomplished by simple gradient descent:

$$\Delta bf_{ij} = \eta < a_i * \left(F_j - \sum_k a_k * bf_{kj} \right) > . \tag{2.7}$$

After each learning step, $\boldsymbol{bf}$ is normalized to ensure that $\sum \|bf\| = 1$. The normalization prevents $\boldsymbol{bf}$ from being unbounded, which would otherwise lead to undesired zero values of $\boldsymbol{a}$.

The sparsely distributed coefficients $\boldsymbol{a}$ then become the output of the encoded layer, which we denote as $\boldsymbol{S}$. A weight matrix between the encoded layer and the hidden layer serves to associate the encoded simple neuron responses related to the same physical stimulus at different retinal positions. Immediately after a saccade takes place, this weight matrix $\boldsymbol{A}$ is updated according to a temporal difference reinforcement learning rule, to strengthen

the weight connections between the neuronal responses to the presaccadic feature to those of the postsaccadic feature.

The neuronal response in the hidden layer $\boldsymbol{H}$ is represented by the following equation:

$$H = A * S. \tag{2.8}$$

The weight matrix $\boldsymbol{A}$ is updated only at those times when a saccade occurs. The updating is done with the following temporal reinforcement learning rule:

$$\Delta A(t) = \eta * [((1-\kappa) * R(t) + k * (\gamma * H(t) - \tilde{H}(t-1))) * \tilde{S}(t-1)], \tag{2.9}$$

where

$$\begin{aligned} \Delta\tilde{H}(t) &= \alpha_1 * (H(t) - \tilde{H}(t-1)) \\ \Delta\tilde{S}(t) &= \alpha_2 * (S(t) - \tilde{S}(t-1)). \end{aligned} \tag{2.10}$$

The factor γ is adjusted to obtain desirable learning dynamics. The parameters η, α_1, and α_2 are learning rates with predefined constant values. In order to investigate the use of the temporal reinforcement in the learning of position invariance, we introduce a weighting parameter κ to balance the importance between the reinforcement reward and the temporal output difference between successive steps. The effect of a varying κ will be demonstrated in section 3.1.

The short-term memory traces, $\tilde{H}$ and $\tilde{S}$, of the neural responses in the hidden layer and the encoded layer are maintained to emphasize the temporal influence of a response pattern at one time step on later time steps. These are temporally low-pass filtered traces of the activities of the hidden layer neurons and encoded layer neurons, respectively. Therefore, the learning rule incorporates a Hebbian term between the input trace and the output trace residuals (the difference between the current and the trace activity), as well as between the input trace and the reinforcement signal. This temporal reinforcement learning rule is not the same as traditional trace rules (Földiák, 1991; Wallis, Rolls, & Földiák, 1993; Wallis & Rolls, 1997), which emphasize the Hebbian connection between the input stimulus and the decaying trace of previous output stimuli.

Equation 2.9 differs slightly from equation 2.1 in Clark and O'Regan (2000), in that we use the temporal difference of the output trace residuals (over longer timescales) instead of the pair-wise temporal difference. This modification enables us to have a longer trace of activities of hidden layer neurons in previous time steps, which helps to obtain more globally optimal solutions.

The reinforcement reward $R(t)$ is the sparsely encoded simple neural response right after a saccade. The weight update rule correlates this reinforcement reward $R(t)$ and (an estimate of) the temporal difference of the

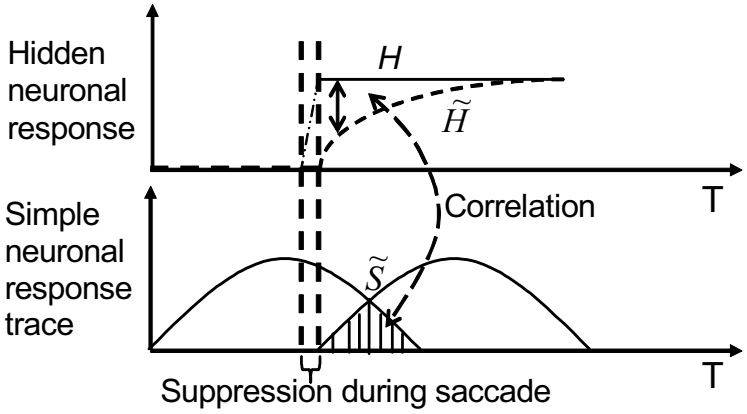


Figure 2: Illustration of temporal difference learning under a temporal perceptual stability constraint. The short-term memory trace of neural response in the encoded layer emphasizes the temporal influence of previous neural responses on later training. The weights between the encoded layer neurons and the hidden layer neurons are enhanced when there is a significant temporal difference between the current hidden neural output and the previous output.

hidden layer neural responses with the memory trace of the encode layer neural responses. The constraint of temporal perceptual stability requires that updating is necessary only when there is a difference between current neural response and previous neural responses kept in the short-term memory trace, as illustrated in Figure 2.

Our proposed position-invariant approach is able to eliminate the limitations of Einhäuser et al.'s model (2002) without imposing an overly strong constraint on the temporal smoothness of the scene images. For example, in the case of recognizing a rapidly moving object, a uniform temporal sampling results in the object appearing in significantly different positions on the retina. This could cause a temporal discontinuity in the input that will cause problems for the Einhäuser et al. model. Even worse, the appearance of the object may have changed due to a change in its pose as it moves. This means that the variation in the input data depends not only on the position of the object, but also on its orientation in space. Such object motion will not affect the learning result of our approach, however, because it employs a nonuniform temporal sampling, in which images are obtained only immediately before and after an attention shift (either overt or covert). As the attention shift takes little time, there is little effect of object motion on the input data. Most of the variation in the position of the object in the image is due to the attention shift.

2.3 Temporal Integration of a Position-Invariant Representation of an Object Across Attention Shifts. The integration level of the executive submodule in our system is concerned with the invariant representation of an object across attention shifts. Position invariance is implicitly incorporated because the attention shift invariance is based on a temporal integration of position-invariant local features.

Attention shift information is provided in our model by the control module. This module receives as input the retinal image of an object (combination of simple features). It constructs a saliency map (Itti, Koch, & Niebur, 1998; Koch & Ullman, 1985) that is used to select the most salient area as the next attention-shift target. The saliency map is a weighted sum of feature saliencies, such as edge orientation, color, and edge contrast. The selection of feature types and their corresponding weights depends on the tasks to be performed. Currently, our implementation uses gray-level images, and we use only orientation contrast and intensity contrast as saliency map features (refer to Itti et al., 1998, for implementation details).

Intensity features, $I(\sigma)$, are obtained from an eight-level gaussian pyramid computed from the raw input intensity, where the scale factor σ ranges from [0..8]. Local orientation information is obtained by convolution with oriented Gabor pyramids $O(\sigma, \theta)$, where $\sigma \in [0..8]$ is the scale and $\theta \in [0^\circ, 45^\circ, 90^\circ, 135^\circ]$ is the preferred orientation.

Feature maps are calculated by a set of "center-surround" operations, denoted by Θ, which are implemented as the difference between fine (at scale $c \in [2, 3, 4]$) and coarse scales (at scale $s = c + \delta$, with $\delta \in [3, 4]$):

$$I(c, s) = |I(c) \Theta I(s)| \tag{2.11}$$

$$O(c, s, \theta) = |O(c, \theta) \Theta O(s, \theta)|. \tag{2.12}$$

In total, 30 feature maps—6 for intensity and 24 for orientation—are calculated and are combined into two conspicuity maps, at the scale ($\sigma = 4$) of the saliency map, through a cross-scale addition $\oplus$:

$$\bar{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(I(c, s)) \tag{2.13}$$

$$\bar{O} = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} N\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(O(c, s, \theta)) \right), \tag{2.14}$$

where $N(\cdot)$ is a map normalization operator.

The saliency map S is obtained by the weighted sum (here, we choose all weights to have the same value) of the two maps:

$$S = \frac{1}{2}(N(\bar{I}) + N(\bar{O})). \tag{2.15}$$

A winner-take-all algorithm (Koch & Ullman, 1985) determines the location of the most salient feature in the calculated saliency map. This location then becomes the target of the next attention shift.

In the case of an overt attention shift, the positional information of the target (including saccadic direction and amplitude) is sent to the executive submodule to command execution of a saccade. The target is foveated after the commanded motion, and a new retinal image is formed. The new image is fed into the module as input for the next learning iteration. A covert attention shift, on the other hand, will not foveate the attended target, and therefore the subsequent retinal image input remains unchanged. Since both overt and covert attention shifts play an important role in determining the timing for learning process at this stage, we use an attention-shift signal instead of a saccade signal as a motor signal from the control module to trigger the integration learning. In the implementation of our model, an inhibition-of-return (IOR) mechanism is added to prevent immediate attention shifts back to the current feature of interest to allow other parts of the object to be explored.

The localized image features, which are obtained when part of an object falls in the attention window before and after attention shifts, are fed into the input layer of the four-layer network. Given that position-invariant representations of local features have already been learned, an integration of local features from an object can be learned in a temporal sequence as long as the attention window stays within the range of the object. Here we assume that attention always stays on the same object during the recognition procedure of an object even in the presence of multiple objects. In our experiments, this assumption is enforced by considering only scenes that contain a single object. In practice, of course, there will be attention shifts between different objects. Although we have not yet tested our method in such situations, it is expected that such interobject attention shifts will only slow learning. This is because a given object will typically be viewed in proximity to a wide range of different objects and backgrounds. Thus, there will be no persistent pairing of an object feature with a particular background feature, and no strong association will be made. The only persistent associations will be those of features within the same object.

The learning at this stage includes two further aspects: a winner-take-all interaction between the output layer neural activities and a fatigue effect on the continuously active output layer neurons. The winner-take-all interaction ensures that only one neuron in the output layer wins the competition to respond actively to a certain input pattern. The fatigue process is a modified implementation of inhibition of return, which prevents one unit from winning all of the input patterns. The fatigue process gradually decreases the fixation of interest on the same object after several attention shifts. Although in our testing we restrict the scenes to contain only one object at a time, we have several objects to be learned on the model. Therefore, it is necessary that the currently active neuron will be suppressed for a while

when the learning moves to the next object. The fatigue effect is controlled by a fixation-of-interest function $\boldsymbol{FOI}(u)$. A u value is kept for each output layer neuron in an activation counter initialized to zero. Each counter traces the recent neural activities of its corresponding output layer neuron. The counter automatically increases by 1 if the corresponding neuron is activated and decreases by 1 until 0 if not. If a neuron is continuously active over a certain period, the possibility of its subsequent activation (i.e., its fixation of interest on the same stimulus) is gradually reduced, allowing other neurons to be activated. A gaussian function of u^2 is used for this purpose:

$$FOI(u) = e^{-u^4/\sigma^2}. \tag{2.16}$$

The output layer neural response $\boldsymbol{C_0}$ is obtained by multiplying the hidden layer neural responses $\boldsymbol{H}$ with the integration weight matrix $\boldsymbol{W}$. $\boldsymbol{C_0}$ is then adjusted by multiplying with $\boldsymbol{FOI}(u)$ and is biased by the local estimation of the maximum output layer neural responses (weighted by a factor $\kappa < 1$):

$$C' = C_0 * FOI(u) - k * \tilde{C}_0. \tag{2.17}$$

If C'_i exceeds a threshold, the corresponding output layer neuron is activated ($\boldsymbol{C}_i = 1$).

The temporal integration of local features is accomplished by dynamically tuning the connection weight matrix between the hidden layer and the output layer. Responses to local features of the same object can be correlated by applying the constraint that output layer neural responses remain constant over time. Given as input the hidden layer neural responses $\boldsymbol{H}$ from the output of the lower layers, and as output the output layer neural responses $\boldsymbol{C}$, the weight matrix $\boldsymbol{W}$ is dynamically tuned in a Hebbian manner using the short-term memory trace $\hat{\boldsymbol{C}}$ of the complex layer neural responses $\boldsymbol{C}$:

$$\Delta W(t) = \gamma * [(\hat{C}(t) - \eta * C(t)) * H(t) - C(t) * W(t)] \tag{2.18}$$

with

$$\Delta \hat{C}(t) = \alpha * (C(t) - \hat{C}(t-1)). \tag{2.19}$$

The short-term memory trace $\hat{\boldsymbol{C}}$ acts as an estimate of the neuron's recent responses. The second term of the learning rule emphasizes the importance of the temporal difference between successive steps in maintaining a stable state. The last term is a local operation that keeps each weight bounded.

2.4 Discussion. The development of the human visual system proceeds gradually from the very basic learning stage, as in the way a newborn baby learns to recognize the complicated external world by exploring simple

shapes and colors step by step. Similarly, in our model, the integration of responses to local features that belong to the same object is based on lower-level extraction of position-invariant local features that have already been learned to some extent. The integration becomes faster when position-invariant representations of local features are correlated in a temporal order rather than the correlation between numerous different neural responses to all local features in random positions. This is also a reason that we do not explicitly distinguish overt and covert attention shift at this stage. In the case of covert attention shifts, although the attended local features are not brought into the fovea, the representations of these peripheral local features are position invariant based on the learning accomplished by the first stage. In the case of overt attention shifts, the attended local features are retargeted to the fovea, and therefore the representations of these local features are already identical to the learned position-invariant representations. Therefore, both types of attention shift can function under this integration.

Our approach is basically a description of a technique for encoding invariant neural responses to changes induced by attention shifts; therefore, attention shifts are an important part of encoding invariant representations for the input patterns, but not necessarily for recognition of an already encoded object. We only need to assume that these attention shifts do occur and that only a single object is being viewed.

For online learning, the two processes of feature extraction and integration are concurrently performed. Because the early learning process of integration is essentially random and has no effect on the later result, we can use a gradually increasing parameter to adjust the learning rate of integration. This parameter can be thought of as an evaluation of the gained experience at the basic learning stage. The value of this parameter is set near 0 at the beginning of the learning and near 1 after a certain amount of learning, at which point the extraction process is deemed to have gained sufficient confidence in its experience on extracting position-invariant local features.

3 Simulation and Results

We designed two experiments to test our model's position-invariant and attention-shift-invariant properties, respectively. In our model, position invariance is achieved when a set of neurons can discriminate one stimulus from others across all positions. We refer to a set of neurons, as our representation is in the form of a population code, in which more than one neuron may exhibit a strong response to one set of stimuli. Between each set of neurons there might be some overlap, but the combinations of actively responding neurons are unique and can therefore be distinguished from each other. We consider attention-shift invariance to have been achieved when the position-invariant set of neurons retains their coherence across attention shifts, when such attention shifts stay on the same object.

We designed a third experiment to show that our model performs better than the models of Földiák (1991) and Einhäuser et al. (2002) when the input patterns lack temporal smoothness.

3.1 Demonstration of Position Invariance. To demonstrate the process of position-invariant local feature extraction, we focus on the extraction submodule. This module is composed of three layers: the input layer, the encoded layer, and the hidden layer. We use two different test sets of local features as training data at this stage: a set of computer-generated images of simple oriented linear features and a set of computer-modified images of real objects.

We first implemented a simplified model that has 648 input layer neurons, 25 encoded layer neurons, and 25 hidden layer neurons for testing with the first training data set. The receptive fields of the input layer neurons are generated by Gabor functions over a 9×9 grid of spatial displacements, each with eight different orientations evenly distributed from 0 to 180 degrees.

The first training image set is obtained by projecting straight lines of four different orientations ([0 degrees, 45 degrees, 90 degrees, 135 degrees]) through a pinhole eye model (as shown in Figure 3) onto seven different positions of a spherical retinal surface. The simulated retinal images each have a size of 25×25 pixels. The training data are shown in Figure 4A, along with a subset of the input layer receptive fields (see Figure 4B).

Figure 5 shows the 25 basis functions (which are the receptive fields of the encoded layer neurons), trained using Olshausen and Field's (1997) sparse coding approach on simple neural responses.

It was found in our experiment that some neurons in the hidden layer responded more actively to one of the stimuli regardless of its positions on the retina than to all other stimuli, as demonstrated in Figure 6. For example, neuron 8 exhibits a higher firing rate to line 4 than to any of the other lines, while neuron 17 responds to line 1 most actively. The other neurons remain inactive to the stimuli, which leaves possible space to respond to other stimuli in the future.

It was next shown that the value of the weighting parameter κ in equation 2.9 had a significant influence on this submodule performance. To evaluate the performance, the standard deviation of activities of the hidden layer neurons are calculated when the submodule is trained with different values of $\kappa (= 0, 0.2, 0.5, 0.7, \text{and } 1)$. The standard deviation of the neural activities is calculated over a set of input stimuli. The value stays low when the neuron tends to maintain a constant response to the temporal sequence of a feature appearing at different positions. Figure 7 shows the standard deviation of the firing rate of the 25 hidden layer neurons with different values of κ. The standard deviation becomes larger as κ increases. This result shows that the reinforcement reward plays an important role in the learning of position invariance. When κ is near 1, which means the learning depends fully on the

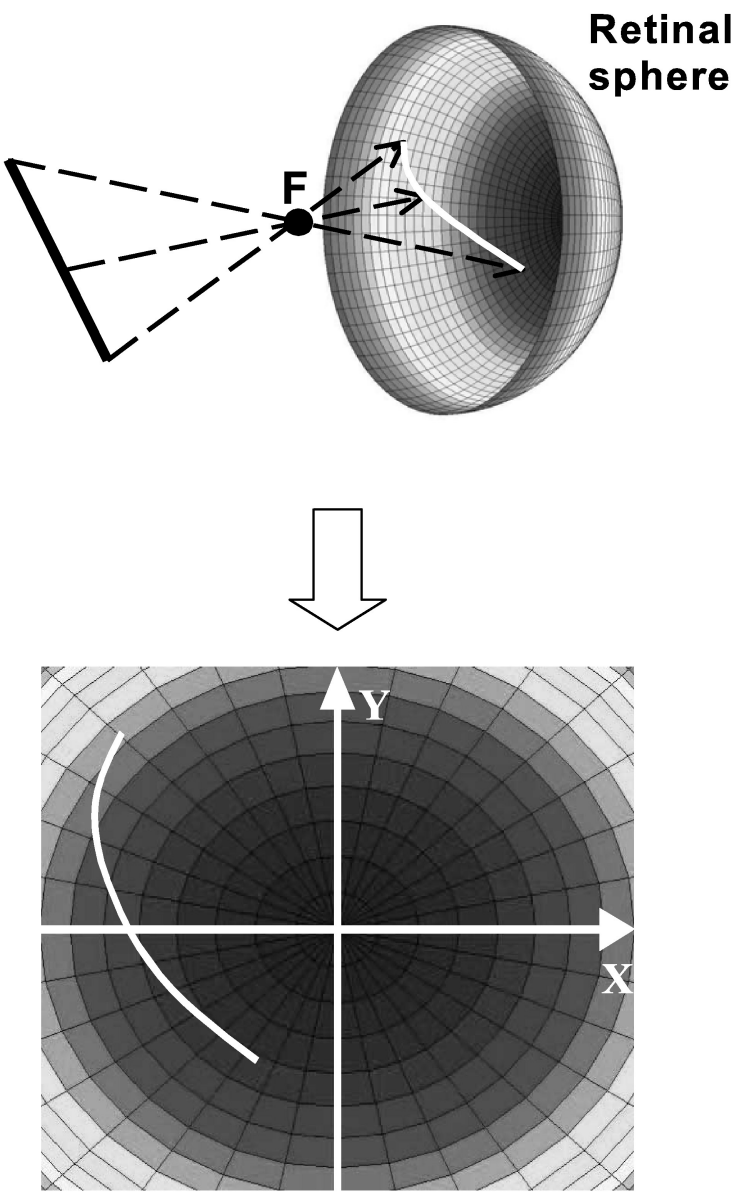
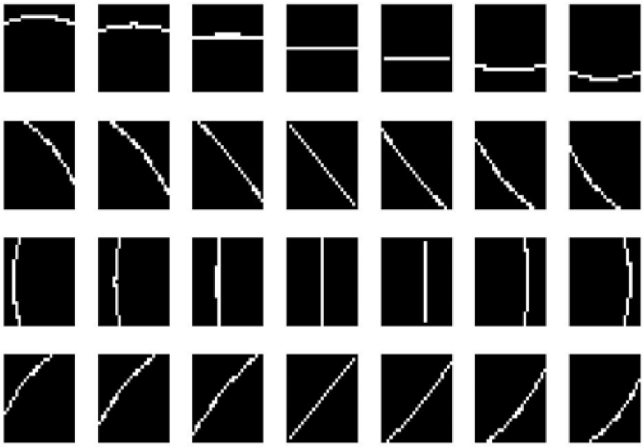
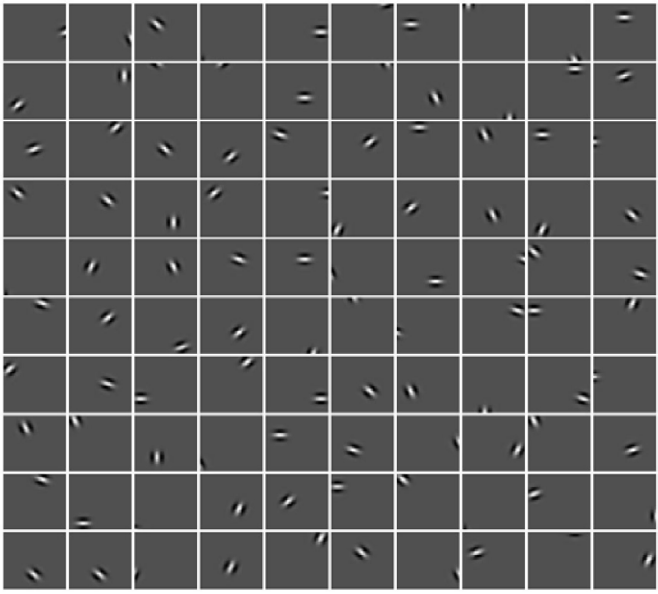


Figure 3: Distorted retinal images obtained when features projected through a pinhole eye model onto the hemispherical surface of the retina.



A



B

Figure 4: (A) Computer-simulated retinal images of lines with four orientations at seven positions used as training data set. (B) A random sample of 100 out of 648 Gabor receptive field profiles of the simple neurons.

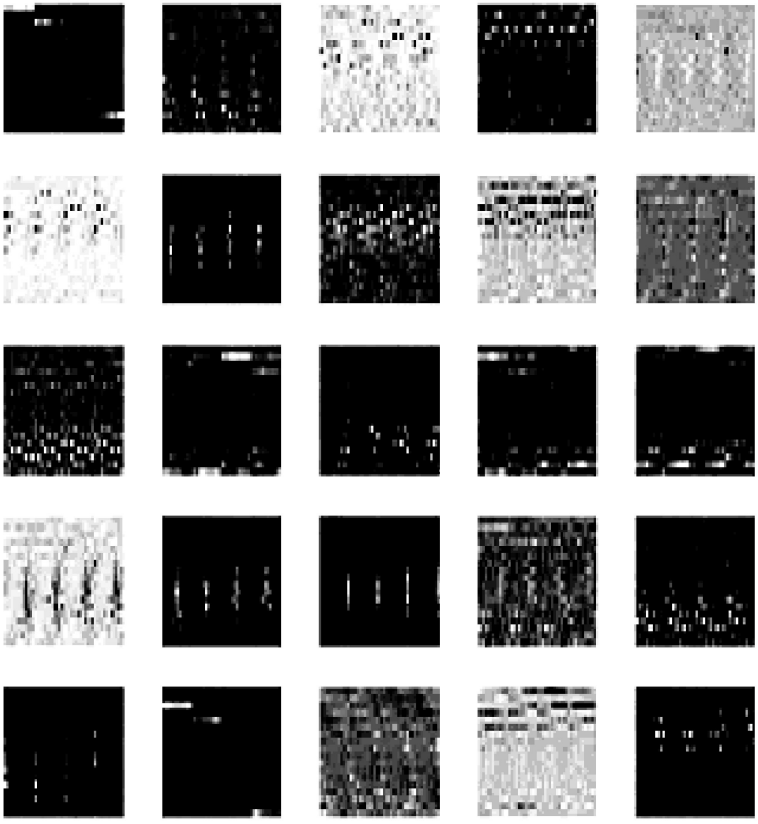


Figure 5: Basis functions visualized as the receptive field profiles of the 25 encoded layer neurons. The basis functions were trained using Olshausen and Field's sparse coding approach.

temporal difference between stimuli before and after a saccade, the hidden layer neurons are more likely to have nonconstant responses.

In our second simulation we tested image sequences of real-world objects, such as a teapot and a bottle (see Figures 8B and 8C). The images of these objects were projected onto the simulated retina at nine different positions following routes such as that illustrated in Figure 8A. Each retinal image has a size of 64×48 pixels. The number of neurons in the encoded layer and the hidden layer has been increased from 25 to 64 from the numbers used in the previous experiment. This was required because the size of the basis function set to encode the sparse representations should also increase as the complexity of the input images increases.

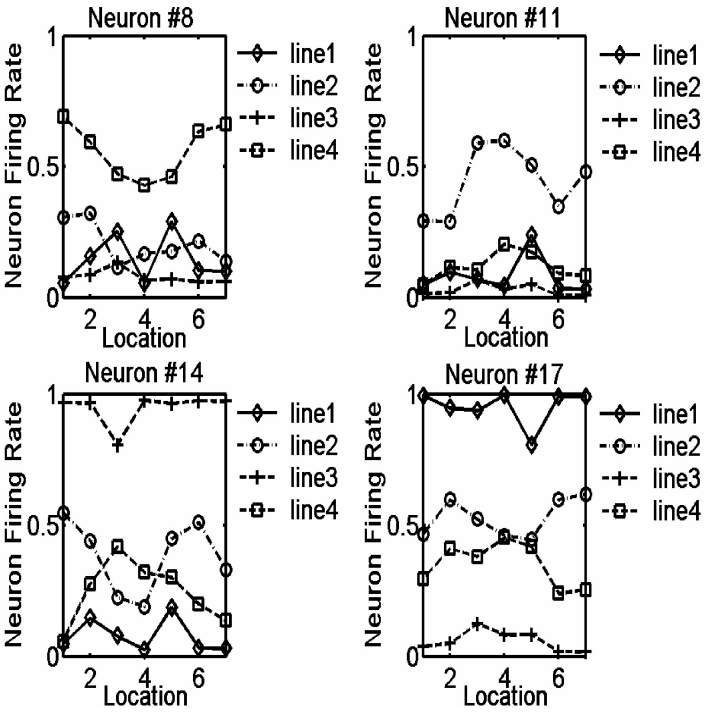


Figure 6: Neural activities of the four most active hidden layer neurons responding to computer-simulated data set at different positions. The neuron firing rates for each of the four stimuli (four lines with different orientations) at each of the seven retinal positions are shown. Each neuron has its preferred orientation selectivity across all positions.

Figure 9 shows the neural activities of the four most active neurons in the hidden layer when responding to the two image sequences of a teapot and a bottle, respectively. Neurons 3 and 54 exhibit relatively strong responses to the teapot across all nine positions, while neuron 27 mainly responds to the bottle. Neuron 25 has strong overlapping neural activities to both stimuli. The sets of neurons that have relatively strong activities are different from each other, satisfying our definition of position invariance.

3.2 Demonstration of Attention Shift Invariance. For simplicity in this experiment, we use binary images of basic geometrical shapes such as rectangles, triangles, and ovals. These geometrical shapes are, as in the previous experiment, projected onto the hemispherical retinal surface through a pin-

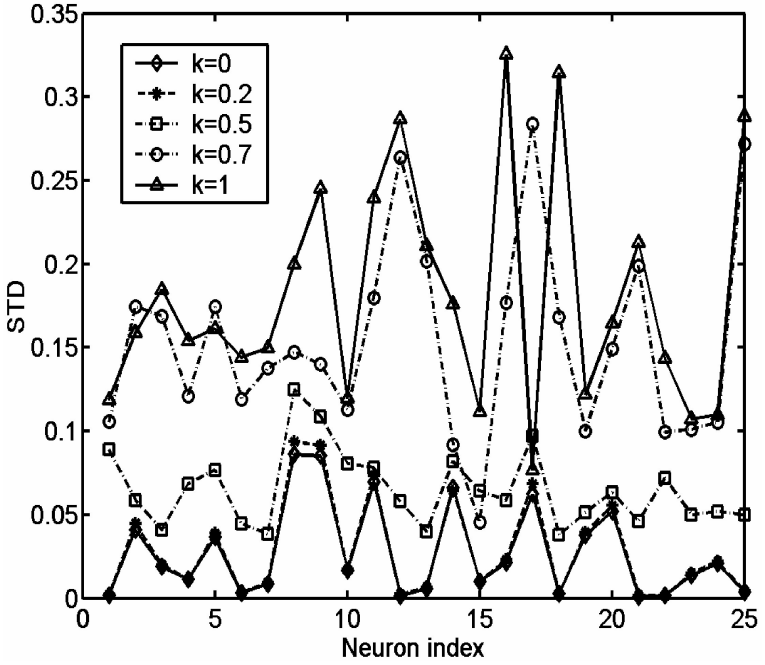


Figure 7: Comparisons of position-invariant submodule performance with varied weighting parameter κ ($\kappa = 0, 0.2, 0.5, 0.7, 1$), using a measurement of standard deviation of each neuronal response to a stimulus across different positions. The weighting parameter emphasizes the importance of the reinforcement reward with small κ. Lower standard deviation values mean that the neural responses remain stable while higher values mean instability. The values for the 25 neurons in the hidden layer are shown.

hole. Their positions relative to the fovea change as a result of saccadic movements.

Here we use a weighted combination of intensity contrast and orientation contrast to compute the saliency map, as they are the most important and distinct attributes of the geometrical shapes we use in the training. A winner-take-all mechanism is employed to select the most salient area as the next fixation target. After a saccade is performed to foveate the fixation target, the saliency map is updated based on the newly formed retinal image, and a new training iteration begins. Figure 10 shows a sequence of saliency maps calculated from retinal images of geometrical shapes for a sequence of saccades.

Figures 11B and 11D show a sequence of pre- and postsaccadic local features of the retinal images of a rectangular shape falling in a 25 $\times$ 25 pixel

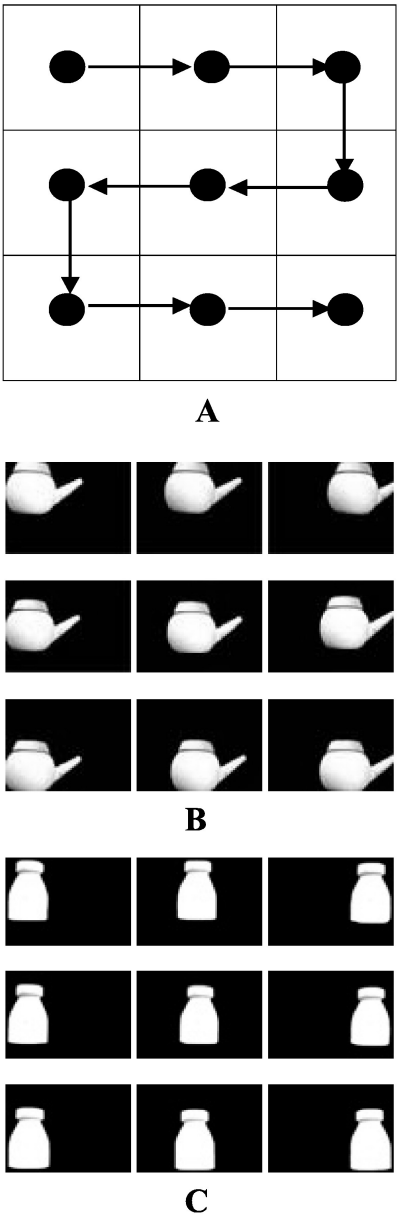


Figure 8: Training image sequences of two real objects (B and C). The images in the sequences were taken at nine positions following a path as indicated in *A*.

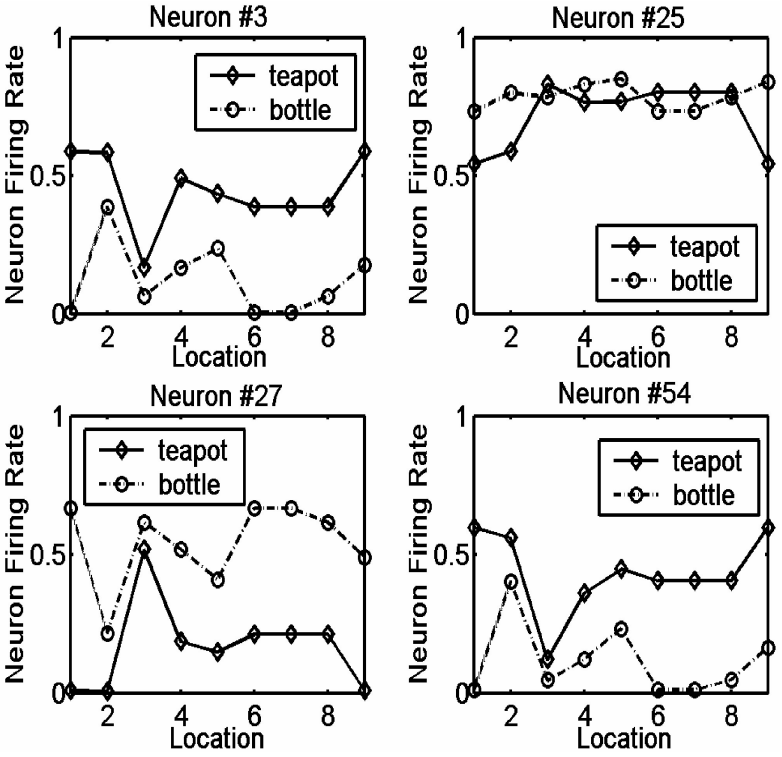


Figure 9: Neural activities of the four most active hidden layer neurons responding to two real objects at different positions. The neuron firing rates for each of the two stimuli (a teapot and a bottle) at each of the nine positions are shown.

attention window, respectively. The local features shown in Figure 11 after a saccade are not exactly the ideal canonical foveal images because of calculation errors in the position of the saccadic target. This situation also occurs in human vision where saccadic eye movements are not always able to put the selected target exactly in the fovea. In fact, undershooting of the target is the usual situation. This undershot local feature is likely to be refoveated by a subsequent small, corrective saccade. An enhanced algorithm dealing with this undershooting was described in Li and Clark (2002). Even if the correction of undershooting is not taken into consideration in this model, we still can obtain invariance, although the efficiency of the model performance will be impaired somewhat. This is because these noncanonical foveal features will exhibit greater variability than the ideal canonical features and therefore require a longer learning process. But the temporal association mechanism is still able to associate the various near-canonical

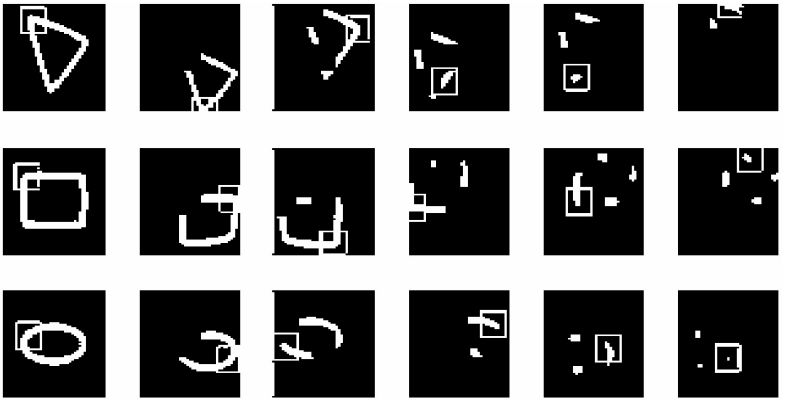


Figure 10: Dynamically changing saliency maps for three geometrical shapes. They are computed from the retinal images after the first six saccades following an overt attention shift. The small, bright rectangle indicates an attention window centered at the most salient point in the saliency map.

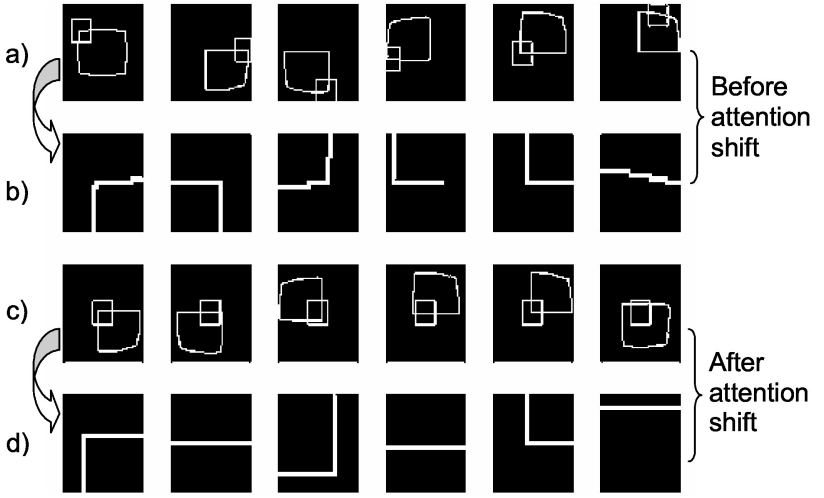


Figure 11: Local features of a rectangular shape before (b) and after (d) an overt attention shift. (a, c) Retinal images of the same rectangle at different positions due to overt attention shifts.

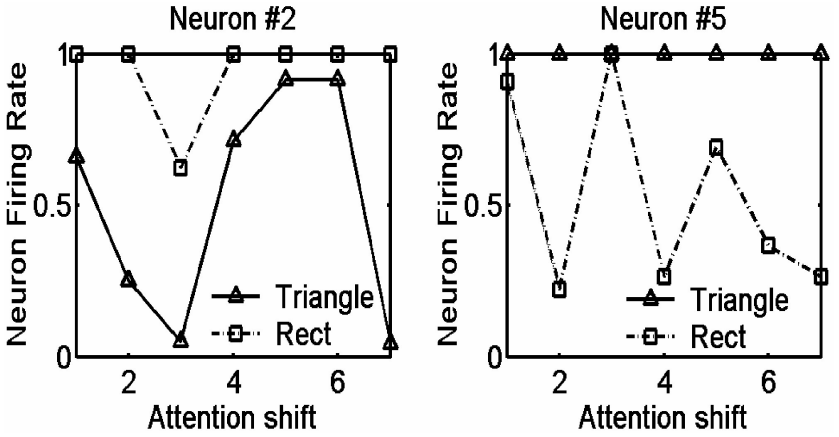


Figure 12: Neural activities of the two most active output layer neurons across attention shifts. The neuron firing rates for the two geometric shapes of a triangle and a rectangle are shown.

neural representations and produce a stable neural response to the same stimulus across transformation.

We show in Figure 12 some of the output layer neural responses (neurons 2 and 5) to two geometrical shapes: a rectangle and a triangle. Neuron 2 responds to the rectangle more actively than to the triangle, while neuron 5 has a more active response to the triangle than to the rectangle.

3.3 Comparison with Other Temporal Approaches. In this section, we demonstrate how our proposed approach performs well in situations where the input lacks smoothness in time. Position-invariant learning models that use temporal continuity, such as those of Földiák (1991) and Einhäuser et al. (2002), are observed to perform poorly in these situations.

We use a digital camera mounted on a computer-controlled pan-tilt unit (PTU) to acquire images around a toy bear. Images of the object are acquired as the PTU randomly changes its pan and tilt positions. The PTU movements are constrained so as to keep the bulk of the object in view at all times. The action of the PTU simulates human eye and head movements, which result

in the displacement of the object features on the imaging surface. Pairs of images before and after each movement are obtained and converted into gray-level images. These image pairs are fed into the model as training data in a random order. The resulting time sequence of images is not smooth at all. This is an extreme test but nonetheless realistic, and it clearly demonstrates the difference in performance between our method and the methods based on temporal continuity.

We implement Földiák's trace rule (1991) and the learning rule for the position-invariant complex neuron of the top layer as given in Einhäuser et al. (2002). We use the training data to train using both of these rules as well as with our proposed model. The learning results are compared using the mean variance of the output neuron responses over the whole stimuli set. If the model is to exhibit position invariance, the output neuron responses should remain nearly constant and therefore have a low variance. We show the results produced by the three models in Figure 13. Each time unit in the plot represents 25 learning iterations. The figure shows that our model converges to a stable state very quickly, with a low mean variance. The mean variance in Einhäuser's model is larger than in our approach, and it descends very slowly over the time interval. Földiák's model produces an increasing response variance with time, implying a complete failure of the learning process for such a nonsmooth input sequence.

4 Conclusions

In this letter, we have presented a neural network model that achieves position invariance. Our approach is based on a study of a more general problem: learning invariance to attention shifts. Attention shifts are the primary reason for images of object features to be projected at various locations on the retina. Object motion in the world is rarely the cause of such variation, as pursuit tracking of object features cancels out this motion. Following Desimone (1990), we treat covert and overt attention shifts as equivalent, from the point of view of their effect on the visual cortex. For the task of learning position invariance, the advantage of treating image feature displacements as being due to attention shifts is the fact that attention shifts are rapid *and* that there is a neural command signal associated with them. The rapidity of the shift means that learning can be concentrated to take place only in the short time interval around the occurrence of the shift. This focusing of the learning solves the problems with time-varying scenery that plagued previous methods, such as those proposed by Földiák (1991), Becker (1993, 1999), Körding and König (2001), and Einhäuser et al. (2002).

We used an extension of Clark and O'Regan's (2000) association model to learn position invariance across overt attention shifts via temporal difference learning on pairs of pre- and postsaccadic stimuli. The extension involves the use of a sparse coding approach, which reduces the size of the association weight matrix and therefore the computational complexity.

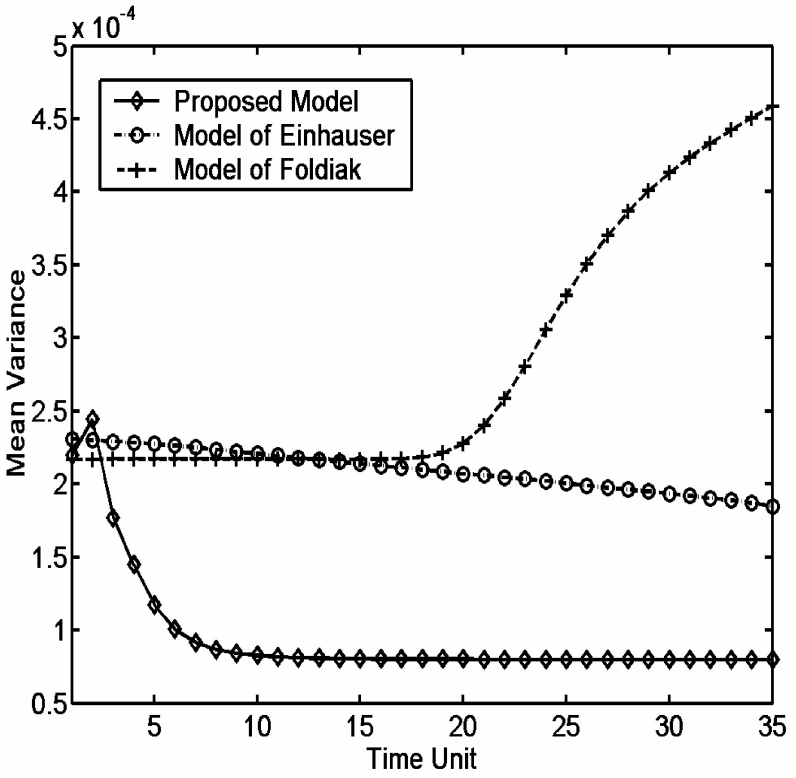


Figure 13: Comparison of the performance of three models (our proposed model, Földiák's, and Einhäuser's respectively) in learning of position invariance in the case of time-varying scenery. The performance is evaluated by mean variance of the hidden layer neuron responses over the whole stimuli set along a time interval. Each time unit in the x-axis is composed of 25 learning iterations.

We apply the constraint of temporal stability across attention shifts, and temporally integrate position-invariant neural response patterns of local features within attention windows to attain attention-shift-invariant object representations.

We implemented a simplified version of our model and tested it with both computer-simulated data and computer-modified images of real objects. In these tests, local features were obtained from retinal images falling in an attention window by an attention shift mechanism. The incorporation of the attention shift mechanism speeds up the learning process by actively acquiring useful information about the correlated relationship between different neural responses of a same local feature at various positions, and

relationship between the partial and the whole (i.e., local features of an object and the object as a whole entity). The results show that our model works well in achieving both position invariance and attention-shift invariance, regardless of retinal distortions. We demonstrated that our method performs well in realistic situations in which the temporal sequence of input data is not smooth, situations in which earlier approaches have had difficulty.

Acknowledgments

We acknowledge funding support from the Institute for Robotics and Intelligent Systems. M.L. thanks Precarn for its financial support.

References

Bartlett, S. M., & Sejnowski, T. J. (1998). Learning viewpoint invariant face representations from visual experience by temporal association. In H. Wechsler, P. J. Phillips, V. Burce, S. Fogelman-Soulie, & T. Huang (Eds.), *Face recognition: From theory to applications* (pp. 381–390). New York: Spinger-Verlag.

Becker, S. (1993). Learning to categorize objects using temporal coherence. In C. L. Giles, S. J. Hanson, & J. D. Cowan (Eds.), *Advances in neural information processing systems, 5* (pp. 361–368). San Mateo, CA: Morgan Kaufmann.

Becker, S. (1999). Implicit learning in 3D object recognition: The importance of temporal context. *Neural Computation, 11*(2), 347–374.

Bridgeman, B., Van der Hejiden, A. H. C., & Velichkovsky, B. M. (1994). A theory of visual stability across saccadic eye movements. *Behavioural and Brain Sciences, 17*(2), 247–292.

Chance, F. S., Nelson, S. B., & Abbott, L. F. (2000). A recurrent network model for the phase invariance of complex cell responses. *Neurocomputing, 32-33*, 339–334.

Clark, J. J., & O'Regan, J. K. (2000). A temporal-difference learning model for perceptual stability in color vision. In *Proceedings of 15th International Conference on Pattern Recognition* (Vol. 2, pp. 503–506). Los Alammitos, CA: IEEE Computer Society.

Desimone, R. (1990). Complexity at the neuronal level (commentary on "Vision and complexity," by J. K. Tsotsos). *Behavioural and Brain Sciences, 13*(3), 446.

Deubel, H., Bridgeman, B., & Schneider, W. X. (1998). Immediate post-saccadic information mediates space constancy. *Vision Research, 38*, 3147–3159.

Einhäuser, W., Kayser, C., König, P., & Körding, K. P. (2002). Learning the invariance properties of complex cells from their responses to natural stimuli. *European Journal of Neuroscience, 15*, 475–486.

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

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*, 193–202.

Gross, C. G., & Mishkin, M. (1977). The neural basis of stimulus equivalence across retinal translation. In S. Harnad, R. Doty, J. Jaynes, L. Goldstein, & G. Krauthamer (Eds.), *Lateralization in the nervous system* (pp. 109–122). New York: Academic Press.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology, 160,* 106–154.

Hyvärinen, A., & Hoyer, P. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research, 41,* 2413–2423.

Ito M., Tamura H., Fujita I., & Tanaka K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology, 73*(1), 218–226.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.

Kikuchi, M., & Fukushima, K. (2001). Invariant pattern recognition with eye movement: A neural network model. *Neurocomputing, 38-40*, 1359–1365.

Koch, C., & Ullman, S. (1985) Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4,* 219–227.

Körding, K. P., & König, P. (2001). Neurons with two sites of synaptic integration learn invariant representations. *Neural Computation, 13*, 2823–2849.

Leopold, D. A., & Logothetis, N. K. (1998). Microsaccades differentially modulate neural activity in the striate and extrastriate visual cortex. *Experimental Brain Research, 123*, 341–345.

Li, M., & Clark, J. J. (2002). *Sensorimotor learning and the development of position invariance*. Poster session presented at the 2002 Neural Information and Coding Workshop, Les Houches, France.

Maunsell, J. H. R., & Cook, E. P. (2002). The role of attention in visual processing. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences, 357*, 1063-1072.

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

Norman, J. (2002). Two visual systems and two theories of perception: An attempt to reconcile the constructivist and ecological approaches. *Behavioural and Brain Sciences, 25*(1), 73–96.

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience, 13*(11), 4700–4719.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*, 607–609.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research, 37*, 3311–3325.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*(5), 939–973.

Perrett, D. I., Rolls, E. T., & Caan, W. (1982). Visual neurons responsive to faces in the monkey temporal cortex. *Experimental Brain Research, 47*, 329–342.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 11*(2), 1019–1025.

Rolls, E. T. (1995). Learning mechanisms in the temporal lobe visual cortex. *Behavioural Brain Research, 66*, 177–185.

Rolls, E. T. (2000). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron, 27*(2), 205–218.

Salinas, M., & Sejnowski, T. J. (2001). Gain modulation in the central nervous system: Where behaviour, neurophysiology, and computation meet. *Neuroscientist, 7*(5), 430–440.

Shewchuk, J. R. (1994). An introduction to the conjugate gradient method without the agonizing pain. Available online at: http://www-2.cs.cmu.edu/~jrs/jrspapers.html.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88*(2), 135–170.

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

Wallis, G., Rolls, E. T., & Földiák, P. (1993). Learning invariant responses to the natural transformations of objects. *International Joint Conference on Neural Networks, 2*, 1087–1090.

Walsh, V., & Kulikowski, J. J. (Eds.). (1998). *Perceptual constancy: Why things look as they do*. Cambridge: Cambridge University Press.

Received April 23, 2003; accepted April 5, 2004.