



Birkbeck ePrints: an open access repository of the research output of Birkbeck College

<http://eprints.bbk.ac.uk>

Spratling, Michael (2005). Learning viewpoint invariant perceptual representations from cluttered images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (5) 753-761.

This is an exact copy of a paper published in *IEEE Transactions on Pattern Analysis and Machine Intelligence* (ISSN 0162-8828). It is reproduced with permission from the publisher. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE. © 2005 IEEE.

Copyright and all rights therein are retained by authors or by other copyright holders. All persons downloading this information are expected to adhere to the terms and constraints invoked by copyright. This document or any part thereof may not be reposted without the explicit permission of the copyright holder.

Citation for this copy:

Spratling, Michael (2005). Learning viewpoint invariant perceptual representations from cluttered images. *London: Birkbeck ePrints*. Available at: <http://eprints.bbk.ac.uk/archive/00000342>

Citation as published:

Spratling, Michael (2005). Learning viewpoint invariant perceptual representations from cluttered images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (5) 753-761.

<http://eprints.bbk.ac.uk>

Contact Birkbeck ePrints at lib-eprints@bbk.ac.uk

Learning Viewpoint Invariant Perceptual Representations from Cluttered Images

Michael W. Spratling

Abstract—In order to perform object recognition, it is necessary to form perceptual representations that are sufficiently specific to distinguish between objects, but that are also sufficiently flexible to generalize across changes in location, rotation, and scale. A standard method for learning perceptual representations that are invariant to viewpoint is to form temporal associations across image sequences showing object transformations. However, this method requires that individual stimuli be presented in isolation and is therefore unlikely to succeed in real-world applications where multiple objects can co-occur in the visual input. This paper proposes a simple modification to the learning method that can overcome this limitation and results in more robust learning of invariant representations.

Index Terms—Computational models of vision, neural nets.

1 INTRODUCTION

INFORMATION about object identity is processed by a series of cortical regions along the ventral pathway leading from the primary visual cortex to the temporal lobe [14], [46]. At each stage along this processing pathway, neurons selectively respond to increasingly complex stimuli. In addition, receptive field sizes also become progressively larger, such that responses are increasingly invariant to stimulus location and scale. For example, cells in area V1 are responsive to simple stimuli, such as oriented edges, at specific locations [17], while neurons in temporal areas are selective to complex objects, such as faces, appearing anywhere in the visual field [6], [21], [28], [29], [32], [42], [45]. Hence, cortical cells along the ventral pathway learn increasing specificity together with increasing invariance [18], [45]. In each cortical region, neurons learn to respond to specific patterns of activity generated by the neurons in more peripheral cortical regions from which they receive their inputs. Higher-level perceptual representations are thus learned from lower-level ones and this process can be repeated hierarchically, such that, at each stage, neurons become tuned to ever more specialized and invariant features of the environment [7], [44]. This process makes tractable the task of learning complex perceptual representations [7] and has formed the basis for many hierarchical neural network models of object recognition.

Different mathematical processes are required for learning more specific representations and for learning more invariant representations. A more specific representation results from a node responding to a combination of coactive lower-level features. A node must thus learn to represent a *conjunction* of presynaptic inputs. In contrast, a more invariant representation results from a node responding to multiple, noncoactive, lower-level features. A node must thus learn to represent a *disjunction* of presynaptic inputs. Hence, several existing

architectures for invariant object recognition (e.g., the Neocognitron [11], [12] and the HMAX model [30], [31]) consist of alternating stages of neural populations that perform these two operations. A simple, two-stage, neural hierarchy of this kind is shown in Fig. 1. It has been proposed [12], [31], [43] that these two forms of processing correspond to the functionality of simple and complex cells observed in the primary visual cortex [16].

What operation is performed by a node critically depends on how its inputs are combined to determine its output (i.e., what *combination function* is employed). To respond to a conjunction of inputs, a standard weighted sum of presynaptic activation values can be used. Such a function will cause the output of the node to be a maximum when all the lower-level features to which it responds are simultaneously active. In contrast, to respond to a disjunction of inputs, a function can be used which causes the output to depend on the maximum input activity [30], [31]. Such a function enables a node to respond invariantly across a number of inputs while maintaining the feature specificity of its response. Hence, an appropriate combination function for responding to a disjunction is obtained by taking the *max* over the inputs, while an appropriate function for representing conjunctions is obtained by taking the *sum* over the inputs.

While many neural networks employ learning rules appropriate to finding conjunctions of inputs, methods for learning disjunctions are not so well established. For example, in the Neocognitron, learning occurs for conjunctions. However, weight-sharing is used to ensure the same feature is learned at different locations and fixed weights are used to pool responses from these nodes to achieve translation invariance [11], [12]. Similarly, in the HMAX model, all the weights are predefined except those that associate the output of the hierarchy with specific object or category representations [30], [31]. Hence, viewpoint invariance is built into both of these architectures, rather than being learned.

To learn a conjunction, it is necessary for a node to form strong connection weights with a set of coactive inputs. By doing so, a neuron learns to become selective for statistical regularities across the input space [1], [9]. To learn a disjunction, it is necessary for a node to form strong connection weights with a set of noncoactive inputs. The

• The author is with the Division of Engineering, King's College, London, UK and the Centre for Brain and Cognitive Development, Birkbeck College, London, UK. E-mail: michael.spratling@kcl.ac.uk.

Manuscript received 18 May 2004; revised 18 Aug. 2004; accepted 11 Oct. 2004; published online 11 Mar. 2005.

Recommended for acceptance by R. Basri.

For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0243-0504.

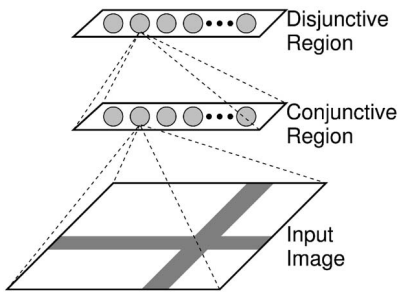


Fig. 1. The proposed neural network architecture for learning perceptual invariances. Nodes in the lower region receive connections from an input image and learn to represent conjunctions of these presynaptic inputs, thus forming more specific representations. Nodes in the upper region receive connections from all the nodes in the lower region and learn to represent disjunctions of these presynaptic inputs, thus forming more invariant representations.

problem is to decide when distinct input patterns result from the same object. "One possible solution to this problem is to associate those images whose appearance is closely temporally correlated, on the assumption that multiple views of an object are frequently experienced in close temporal succession" [51]. By doing so, a neuron learns to become selective for statistical regularities across time. This suggestion [15] has formed the basis for a large number of algorithms that learn invariance from sequences of images, e.g., [2], [3], [4], [5], [8], [10], [19], [24], [25], [26], [33], [38], [39], [40], [43], [47], [48], [49], [54], [55], [56]. This approach is justified on the grounds that objects are seen for periods of time, during which they may undergo a number of transformations or be observed from a number of viewpoints. Furthermore, both psychophysical and physiological evidence indicates that object representations are learned from temporal associations [22], [34], [37], [41], [50], [51], [53]. One popular implementation of this form of learning modifies an activity-dependent learning rule such that a moving average (or short-term *trace*) of previous activity is used instead of a value or instantaneous activity. This enables previous activity to influence learning of subsequent input patterns and, hence, transforms temporal regularities into spatial ones [47].

Despite the large number of different algorithms that have been developed to exploit temporal correlations in learning invariant representations, none have been proposed that can work in more realistic environments that may contain multiple objects or background clutter. Previous algorithms thus require stimuli to be presented in isolation [10], [24], [40], [48] and, hence, would be inappropriate for real-world applications. The following section describes a simple modification to the learning method that can overcome this limitation. The proposed learning rule cannot only successfully function when multiple objects co-occur, but its performance is actually enhanced in such circumstances. The proposed algorithm thus provides a more robust and more efficient mechanism for learning invariance.

2 METHOD

When multiple objects are present in a scene, learning invariance purely by finding temporal associations across image sequences would require knowledge of which representations derived from each image corresponded to the same object. This is a temporal version of the correspondence

problem that occurs in stereo vision. Solving this complex binding problem for every pair of images in a sequence would be computationally expensive. However, it is also possible to employ another constraint in the learning process in order to learn invariant representations without solving the correspondence problem. In terms of translation, this constraint exploits the fact that the same object cannot occupy two distinct locations at once. However, the proposed method is not limited to learning invariance under translation, but is applicable to any form of transformation. In more general terms, the proposed additional constraint exploits the fact that the same object cannot generate two distinct percepts at once. Since the proposed method takes advantage of an additional source of information within the image data, learning of invariance is actually improved by having multiple objects appear simultaneously.

Current methods for learning invariant representations using a short-term trace of previous activity cause synaptic weights to be increased when there is correlation between previous outputs and the current inputs (or between the current outputs and previous inputs, depending on the implementation details). In contrast, the proposed method allows such synaptic weight increases for only one input and decreases the weights from all other simultaneously active inputs (implemented by (6)). Hence, when multiple object representations are active in response to a single image, the proposed learning mechanism biases disjunctive learning in the next stage of the hierarchy so that only one of these representations is encoded by any individual node.

In addition, the proposed algorithm does not employ a trace of previous activity, but learns correlations between the current input and the output generated in response to the previous image. Hence, temporal associations are learned between consecutive pairs of images, rather than across longer sequences of images. With this proposed mechanism, the output generated by the current image does not affect learning at the current iteration (it will only affect learning at the next iteration). This was also the case for the learning rule employed by Rolls and Milward [33], who found that learning was improved when a trace of previous activity was used that did not include a contribution from the current iteration. However, the proposed method not only excludes the activity in response to the current image from affecting learning, but it also excludes effects from all activity generated prior to the previous image and, hence, eliminates the trace completely. This has the advantage that there is no need to calculate or remember a trace of previous activity. Furthermore, there is no parameter controlling the length of the trace and, hence, no requirement to adjust the trace time constant to be appropriate for different tasks [24], [33], [55].

Several previous models which learn invariance have calculated node activity using a weighted sum of the inputs, e.g., [10], [24]. As mentioned in Section 1, the \max operator has been proposed as a more appropriate combination function for generating invariant responses [30], [31]. Summation allows multiple, nonoptimal, stimuli to generate as strong a response as the optimal stimulus. For example, rather than selectively responding to a particular object, a node employing summation might be fully activated by disjoint object parts occurring at a range of locations. Furthermore, a node employing summation would be more active in response to multiple occurrences of the same object, than to a single object. To avoid these problems, the algorithm

proposed here employs the \max operator for finding disjunctions (see (1)). The proposed model is thus distinct in that it both learns invariance and employs the \max operator. In contrast, previous models that employ an appropriate combination function for generating invariant responses fail to learn those invariances, e.g., [30], [31], while those models that do learn invariances use an inappropriate combination function, e.g., [10], [24].

2.1 Implementation Details

The performance of the proposed algorithm was explored using a simple neural hierarchy consisting of two stages, or regions (see Fig. 1). One region learned conjunctions and the other region learned disjunctions. The conjunctive region was implemented using the algorithm presented in [35], [36], but with minor modifications which improve performance. Implementation details for both types of region are given below.

2.1.1 Activation

As discussed above, different combination functions are appropriate for learning disjunctions and conjunctions. Hence, in the proposed algorithm, the activations of nodes in the disjunctive region were calculated using the \max operator:

$$y_{jk}^t = \max_{i=1}^m \{Z_{ijk}^t\} \quad (1)$$

and the activations of nodes in the conjunctive region were calculated using summation:

$$y_{jk}^t = \sum_{i=1}^m (w_{ijk} X_{ijk}^t), \quad (2)$$

where y_{jk}^t is the activation of node j in region k at time t , m is the total number of inputs to the region, w_{ijk} is the synaptic weight from input i to node j in region k , Z_{ijk}^t is the normalized weighted activation received at input i of node j in region k , and X_{ijk}^t is the activation received by node j in region k from input i after preintegration lateral inhibition:

$$Z_{ijk}^t = x_{ijk}^t \left(\frac{w_{ijk}}{\max_{q=1}^m \{w_{qjk}\}} \frac{w_{ijk}}{\max_{q=1}^n \{w_{iqk}\}} \right), \quad (3)$$

$$X_{ijk}^t = x_{ijk}^t \left(1 - \alpha^t \max_{\substack{p=1 \\ (p \neq j)}}^n \left\{ \frac{w_{ipk}}{\max_{q=1}^m \{w_{qp k}\}} \frac{y_{pk}^{t-1}}{\max_{q=1}^n \{y_{qk}^{t-1}\}} \right\} \right)^+, \quad (4)$$

where x_{ijk}^t is the input activity received at input i of node j in region k , y_{pk}^{t-1} is the activation of node p in region k at time $t-1$, n is the total number of nodes in the region, α^t is a scale factor controlling the strength of lateral inhibition, and $(v)^+$ is the positive half-rectified value of v . The presynaptic activity values (x_{ijk}^t) are either the activations of nodes in the lower region at the previous time step (i.e., y_{jk}^{t-1}) or are external, sensory, inputs supplied to the hierarchy.

The value of Z_{ijk}^t depends upon the strength of presynaptic activity and the strength of the weight received from that input (i). This value is adjusted using both post and

presynaptic weight normalization. Such normalization causes the value of Z_{ijk}^t to be reduced if the node receives a stronger connection from another input or if another node receives a stronger connection from input i . The value of Z_{ijk}^t is thus dependent on prior weight changes that have taken place in this and other nodes. The first form of normalization biases the node to respond to an active input to which it has previously responded, while the second form of normalization biases the network to respond to each input using a single node. Through activity-dependent learning (see Section 2.1.2), nodes thus become selective for disjunctive sets of input patterns and each disjunctive set is represented by an individual node.

The value of X_{ijk}^t depends upon the strength of presynaptic activity and the strength of the lateral inhibition directed toward that particular input. The value of X_{ijk}^t will be strongly inhibited if another node (p) is strongly activated by the overall stimulus (i.e., if y_{pk}^{t-1} has a high value relative to all other node activations) and that other node receives a strong synaptic weight from input i (i.e., if w_{ipk} has a high value relative to all the other weights received by node p). Hence, this form of lateral inhibition provides competition by enabling each node to “block” its preferred inputs from activating other nodes. There is thus strong competition which causes nodes to become selective for distinct conjunctive sets of inputs, but which does not prevent multiple nodes from responding simultaneously to the presentation of multiple, distinct, stimuli. To help ensure that a steady-state solution is reached, it has been found useful to gradually increase the value of α^t at each iteration from an initial value of zero. In the simulations described in this paper, the value of α^t was increased from zero to 10 in steps of 0.25, while the input image was kept fixed. Activation values generally reached a steady-state at lower alpha, in which case, the competition was terminated prior to α^t reaching its maximum value.

Finally, the activity of each node was also modified by a small amount of noise such that:

$$y_{jk}^t = y_{jk}^t (1 + \rho). \quad (5)$$

The noise values, ρ , were logarithmically distributed positive real numbers in the range $[0, 0.01]$. This noise is essential to cause nodes to learn to represent distinct stimuli. Since the magnitude of the noise is small, it has very little effect on neural activity except when multiple nodes have virtually identical synaptic weights. When this occurs, the noise causes one of these nodes to win the competition to be active in response to the current stimulus.

2.1.2 Learning

All synaptic weights were initially given equal values. Weights were modified using the final, steady-state, activation values found using the equations given above (the t superscript is thus dropped from subsequent equations). For nodes in a disjunctive region, the following learning rule was employed:

$$\Delta w_{ijk} = \pm \frac{\gamma x_{ijk}}{\sum_{p=1}^n y_{pk}^*} (y_{jk}^* - \bar{y}_k^*)^+, \quad (6)$$

where γ is a parameter controlling the learning rate ($\gamma = \frac{1}{4}$ was used in the simulations presented in this paper), y_{jk}^* was the activity of the node in response to the previous input

pattern, and \bar{y}_k^* was the mean of the output activations in response to the previous input pattern (i.e., $\bar{y}_k^* = \frac{1}{n} \sum_{j=1}^n y_{jk}^*$). Using y_{jk}^* causes learning of the correlation between the current input activity and the previous output activity. Learning only occurs for nodes that were more active than average at the previous iteration and at synapses with currently active inputs. This learning rule has a positive value at synapses where $Z_{ijk} = y_{jk}$ and a negative value otherwise. Hence, only the weight of the most active input was increased and weights to all other active inputs were decreased. This learning rule thus encourages each node to learn weights selective for a set of noncoactive inputs. This is achieved since, when a node is more active than average, it increases its synaptic weights to a single active input and decreases its weights to all other active inputs. Hence, only sets of inputs which are seldom coactive will generate strong afferent weights. This rule thus enforces the proposed additional constraint for learning disjunctions: that the same object cannot generate two distinct percepts (i.e., inputs) at the same time. Following learning, synaptic weights were clipped at zero such that $w_{ijk} = (w_{ijk})^+$ and were normalized such that $\sum_{j=1}^n w_{ijk} = 1$. This normalization process provides an implicit form of competition between different disjunctive nodes since, if one node strengthens its connection to a particular input, then connections from that input to all other nodes are weakened.

For nodes in a conjunctive region, the following learning rule was employed for weights with values greater than or equal to zero:

$$\Delta w_{ijk} = \frac{\beta(x_{ijk} - \bar{x}_{jk})}{\sum_{p=1}^m x_{pjk}} (y_{jk} - \bar{y}_k)^+, \quad (7)$$

where β is a parameter controlling the learning rate ($\beta = 1$ was used in the simulations presented in this paper), \bar{x}_{jk} is the mean of the input activations (i.e., $\bar{x}_{jk} = \frac{1}{m} \sum_{i=1}^m x_{ijk}$), and \bar{y}_k is the mean of the output activations (i.e., $\bar{y}_k = \frac{1}{n} \sum_{j=1}^n y_{jk}$). Following learning, synaptic weights were clipped at zero such that $w_{ijk} = (w_{ijk})^+$ and were normalized such that $\sum_{i=1}^m (w_{ijk})^+ = 1$. This learning rule encourages each node to learn weights selective for a set of coactive inputs. This is achieved since, when a node is more active than average, it increases its synaptic weights to active inputs and decreases its weights to inactive inputs. Hence, only sets of inputs which are consistently coactive will generate strong afferent weights. In addition, the learning rule is designed to ensure that nodes can represent stimuli which share input features in common (i.e., to allow the network to represent overlapping patterns). This is achieved by rectifying the postsynaptic term of the rule so that no weight changes occur when the node is less active than average. If learning was not restricted in this way, whenever a pattern was presented, all nodes which represented patterns with overlapping features would reduce their weights to these features.

For weights with values less than or equal to zero, the following learning rule was employed:

$$\Delta w_{ijk} = \frac{\beta(X_{ijk} - 0.5x_{ijk})^-}{\sum_{p=1}^n y_{pjk}} (y_{jk} - \bar{y}_k), \quad (8)$$

where X_{ijk} is the input activation from source i to node j after inhibition (see (4)), and $(v)^-$ is the negative half-rectified value of v . Negative weights were constrained such that

$0 \geq \sum_{i=1}^m (w_{ijk})^- \geq -1$. This learning rule is only applied to synapses which have a weight of zero (or less than zero) caused by application of the learning rule given in (7) (or prior application of (8)). Negative weights are generated when a node is active and inputs that are not part of the nodes' preferred stimulus, are inhibited. This can only occur when multiple nodes are coactive. If the pattern to which this set of coactive nodes are responding reoccurs, then the negative weights will grow. When the negative weights are sufficiently large, the response of these nodes to this particular pattern will be inhibited, enabling other nodes to successfully compete to represent this pattern. On the other hand, if the pattern to which this set of coactive nodes are responding is just due to the coactivation of independent input patterns, then the weights will return toward zero on subsequent presentations of these patterns in isolation.

2.1.3 Comparison with Previous Methods

In contrast to previous methods, the above algorithm for learning temporal correlations employs the activity at the previous iteration rather than a trace of previous activity. To determine the effects of this, certain experiments were repeated, using an identical procedure to that described above, but with the y_{jk}^* value in (6) replaced by a trace of previous activity such that:

$$y_{jk}^* = 0.2y_{jk} + 0.8y_{jk}^*.$$

An identical equation for calculating a trace of previous activity has been employed in several papers, i.e., [10], [47], [48].

To compare the results of the proposed algorithm with previous methods, certain experiments were also repeated, using the algorithm proposed in [10] for the disjunctive region. Hence, (6) was replaced by:

$$\Delta w_{ijk} = 0.02(x_{ijk} - w_{ijk})y_{jk}^*,$$

where y_{jk}^* was a trace of previous activity, calculated as above, and (1) was replaced by:

$$y_{jk}^t = \sum_{i=1}^m (w_{ijk} x_{ijk}^t),$$

following which a winner-takes-all competition was applied so that the node receiving the strongest weighted input was given an activity of one, and all other nodes were given an activity of zero. This algorithm has formed the basis for many previous methods of learning invariance, e.g., [24], [40], [47], [48] and, so, will be referred to as the standard method.

3 RESULTS

The proposed algorithm was applied to the task of learning line orientation invariant to translation. A similar task was used previously to test other methods [10], [19], [24]. A two-region hierarchy of neural networks was used. The lower region received input from images of training stimuli and learned conjunctions of features within these images, while the upper region received input from all the nodes in the lower region and learned disjunctions across the activity patterns generated in the lower region. A network consisting of 32 nodes was used for the lower region and a network containing five nodes was employed for the upper

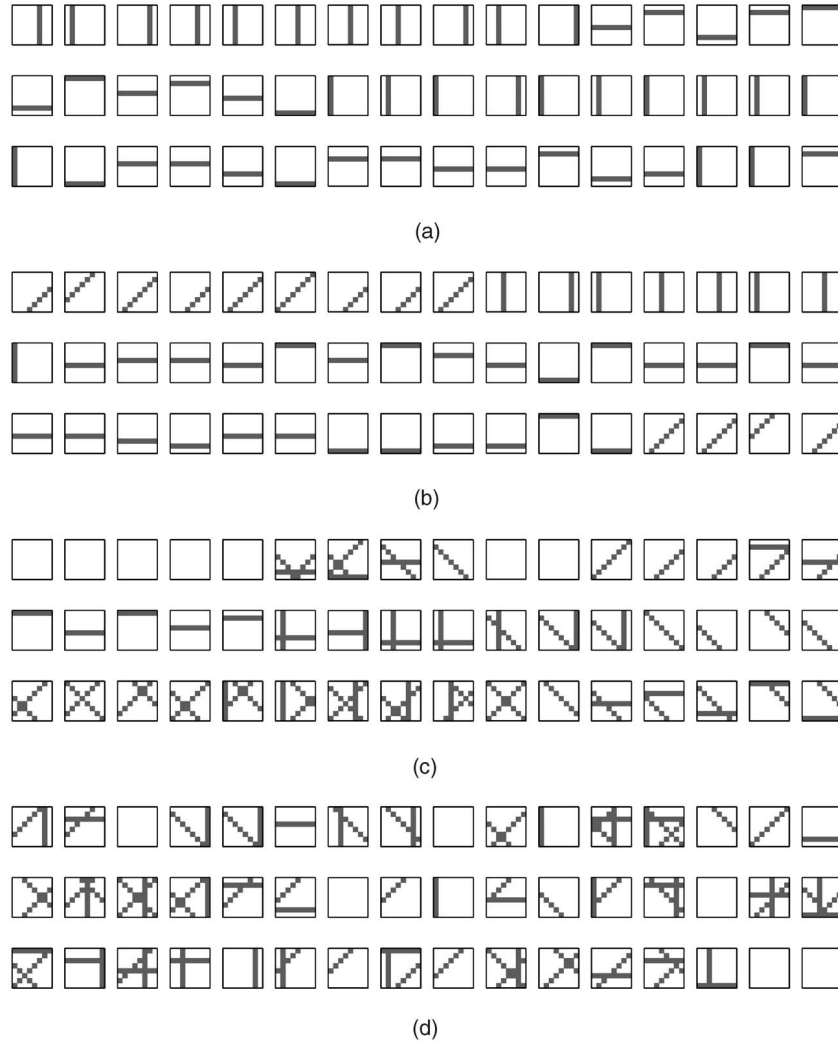


Fig. 2. Typical examples of image sequences used for learning bar orientations invariant to translation. Bars were selected such that orientations were mutually exclusive in (a) and (b) and bar orientations were independently selected in (c) and (d). Images could contain bars at two orientations (horizontal and vertical) in (a) and four orientations (horizontal, vertical, and both diagonals) in (b), (c), and (d). The probability that successive images contained bars of the same orientation was 0.9 in (a), (b), and (c) and 0.6 in (d).

region. Learning in both networks proceeded from the start of training (i.e., there was no pretraining of the lower network before training the upper network).

Training data consisted of a series eight-by-eight pixel images, within which one-pixel wide horizontal, vertical, and diagonal bars could appear. Two methods for generating the training data were used. Bars could be selected such that orientations were mutually exclusive or the bar orientations present in each image could be independently selected:

Mutually exclusive orientations: Each image contained only one bar at a time. At each iteration, the orientation of this bar could be changed with a fixed probability.

Independently selected orientations: Each image could contain an arbitrary number of bars at different orientations. At each iteration, each orientation could be changed from being present to absent (or vice versa) with a fixed probability.

In both cases, specific bars at the chosen orientation(s), were selected at random. The algorithm was tested using a range of values for the probability of successive images containing bars of the same orientation and with either two or

four different orientations being used (i.e., images could contain only horizontal and vertical bars in some experiments, but horizontal, vertical, and diagonal bars¹ in others). Fig. 2 shows typical examples of the type of image sequences that were used.

To succeed in the task, nodes in the lower region needed to learn to represent all the individual bars at all locations and distinct sets of nodes in the upper region needed to learn weights such that they received the strongest connections projected by all the nodes in the lower region which represented bars of a single orientation. Fig. 3 shows a typical example of the weights learned when the hierarchy succeeds in learning to represent horizontal and vertical bars. For each combination of orientation selection method, number of bar orientations, and probability of successive images containing bars of the same orientation, the hierarchy was trained 10 times using different randomly generated sequences of input patterns. The number of these trials for which the task was successfully learned are shown in Fig. 4. For trials in which two orientations were present, the hierarchy was trained for

1. The seven longest diagonals in each direction.

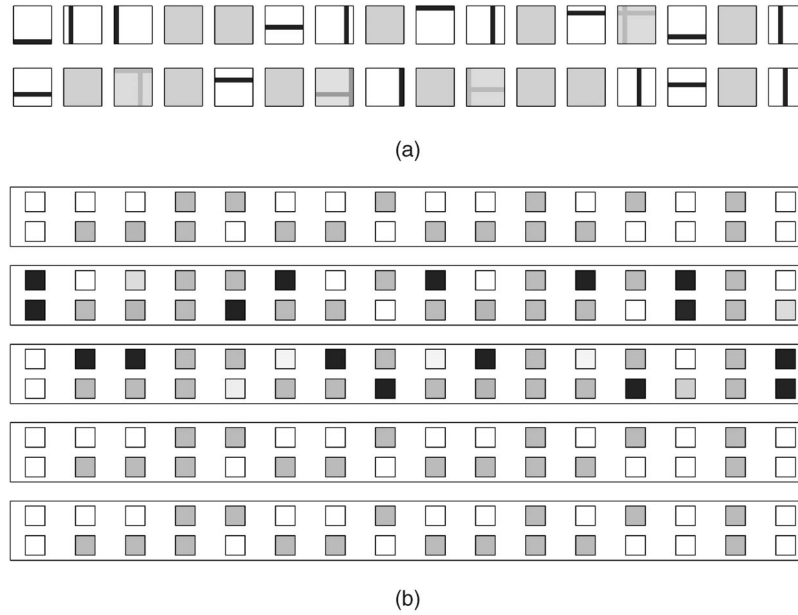


Fig. 3. A typical example of the synaptic weights formed by learning to represent horizontal and vertical bars. A 5-node disjunctive network received input from a 32-node conjunctive network which, in turn, received input from an 8 by 8 pixel input array. The synaptic weights, learned after 5,000 iterations, for (a) the 32 nodes in the lower network and (b) the 5 nodes in the upper network. For each node in the upper network, the weights received from each of the 32 nodes in the lower network are shown in the same order in which these nodes are presented in (a). The darkness of each pixel is proportional to the strength of that weight. Nodes in the lower network have become selective for sets of pixels which correspond to individual bars at each location. One node in the upper network has learned to represent horizontal bars, while another has learned to represent vertical bars: Each of these nodes is selective for all the nodes in the lower network which represent bars of that orientation.

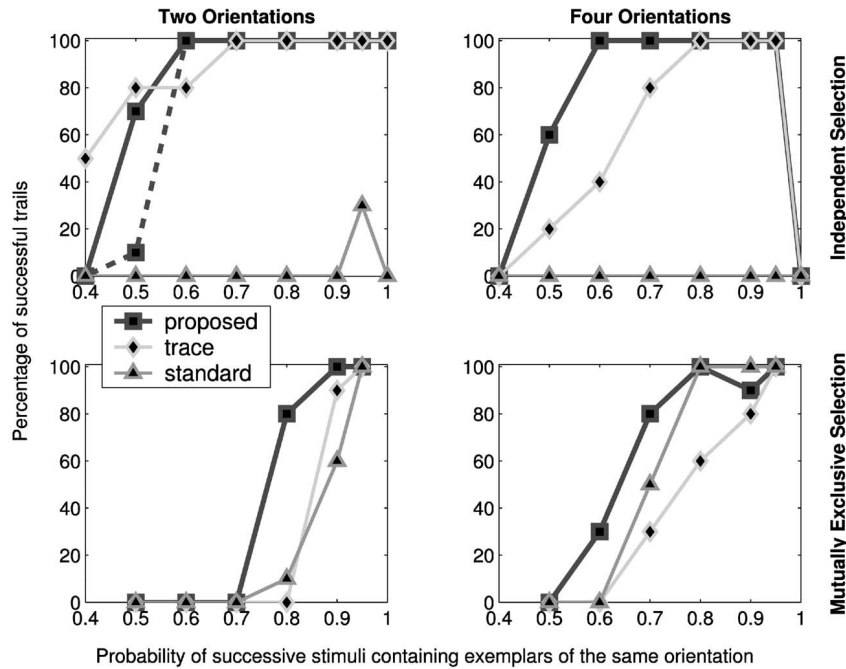


Fig. 4. Reliability of learning representations of line orientation invariant to translation. Each graph shows the percentage of trails for which the hierarchy successfully learned distinct representations for bars of each orientation at all image locations. Within each graph, the number of successful trials is plotted for varying the probability that successive images will contain bars of the same orientation. The left column shows results when two orientations of bars (horizontal and vertical) were used and the right column shows results when four orientations of bars (horizontal, vertical, and both diagonals) were used. The top row shows results when bar orientations were independently selected (i.e., when multiple bars could be present within the same image) and the bottom row shows results when bar orientations were mutually exclusive (i.e., each image contained only one bar). Plots are shown for the proposed algorithm, for the proposed algorithm using a trace of previous activity, and for the standard algorithm. The dashed line shows the performance of the proposed algorithm when 10 percent of images contained two bars of the same orientation.

5,000 iterations and, for trials in which four orientations were used, since there were twice as many representations to be learned, the hierarchy was trained for 10,000 iterations.

It can be seen that the proposed algorithm reliably learns to represent bar orientation, invariant to position, across a wide range of conditions. As with previous methods, learning

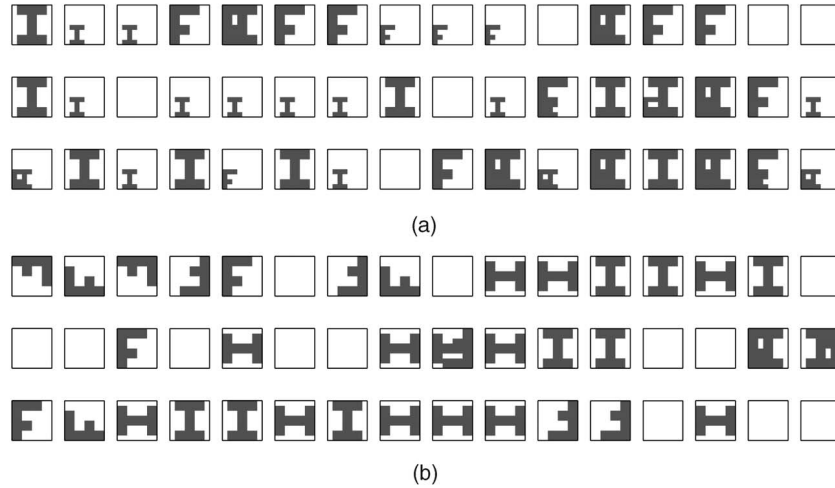


Fig. 5. Typical examples of image sequences used for learning invariance to scale and rotation. Images containing the letters “F” and “I” at (a) different scales and (b) different orientations were created by independently selecting exemplars from each class to be present. In these examples, the probability that successive images contained the same letter was 0.7.

relies on different views of the same object being presented across sequences of images. Hence, decreasing the probability that successive images contain bars of the same orientation (i.e., shortening the average length of image sequences containing bars of the same orientation) leads to a reduction in performance and, eventually, to failure to learn the task in all trials. The point at which the algorithm fails to learn reliably, occurs at shorter sequence lengths when images contain multiple rather than isolated bars (compare the top and bottom rows of Fig. 4). Hence, the presence of multiple objects within the visual input improves learning using the proposed method. This is to be expected since the proposed algorithm can exploit an additional source of information when multiple objects appear simultaneously. Note that all bar orientations were present in every image for the condition in which independent selection was used together with a probability of one that the same orientation is present in successive images. The failure to successfully learn in this condition, when using four bar orientations, resulted from the lower-region failing to represent all the bars in nine out of 10 trials. It is thus primarily a failure of conjunctive rather than disjunctive learning. Hence, when the probability that successive images contain bars of the same orientation is very high, the lower (conjunctive) network can fail to learn successfully, while, at low probabilities, the upper (disjunctive) network fails.

The proposed learning mechanism assumes that, when multiple object representations are active in response to a single image, these representations cannot correspond to different views of the same object. This assumption will be invalid in rare situations when two, or more, identical objects are present in the same scene. To ensure that the proposed method is still robust even in such circumstances, experiments were repeated so that 10 percent of images would contain two bars of the same orientation. Results are shown as a dashed line in the top left-hand plot of Fig. 4. It can be seen that, when images may contain multiple views of the same object, a reduction in the number of successful trials only occurs for very short sequence lengths. For longer sequence lengths, learning is still reliable.

The proposed algorithm learns correlations between the current input and the output generated in response to the

previous image. When the proposed learning rule was modified to use a trace of previous output activity, results were worse in all conditions where images contained only single bars. When trained using isolated bars, the proposed algorithm using a trace learning rule produces similar results to the standard method of learning invariance (which also uses a trace). When training images contained multiple bar orientations, employing a trace of previous output activity with the proposed algorithm also generally resulted in worse performance. However, when only two bar orientations were used and bars of the same orientation were presented for very short sequence lengths, employing a trace in the proposed learning rule appears to improve performance. However, in these conditions, the trace does not provide any useful information and replacing y_{jk}^* in (6) with a value of $1 + \rho$ (where ρ is a random variable in the range $[0, 0.01]$) produced even better results (i.e., a 100 percent success rate for both two and four orientations and a probability of 0.4 that successive stimuli contain exemplars of the same orientation). Hence, it appears that, for very short sequence lengths, excluding node output activity from making any contribution to learning enables the constraint that coactive stimuli should not be represented by the same node to be even more successful in detecting distinct bar orientations. In contrast, the standard method completely fails when images contain multiple objects. The standard method requires objects to be presented in isolation and the proposed method is therefore far more robust in situations where this does not occur.

The above results provide examples of learning invariance to translation. However, the proposed method can also be applied to learning invariant perceptual representations under other forms of transformation. An upper region node learns to respond to a disjunctive set of lower region nodes. Hence, if the lower region nodes learn to represent distinct views of an object under arbitrary transformations, then the upper region can learn to represent all these different views of the same object. As a simple illustration of this point, the neural hierarchy described above was trained using eight-by-eight pixel images, within which the letters “F” and “I” could appear, either at two different scales or at different in-plane rotations (see Fig. 5 for typical examples of the image sequences that were used). For each of a range of probabilities that successive images contained exemplars of the same

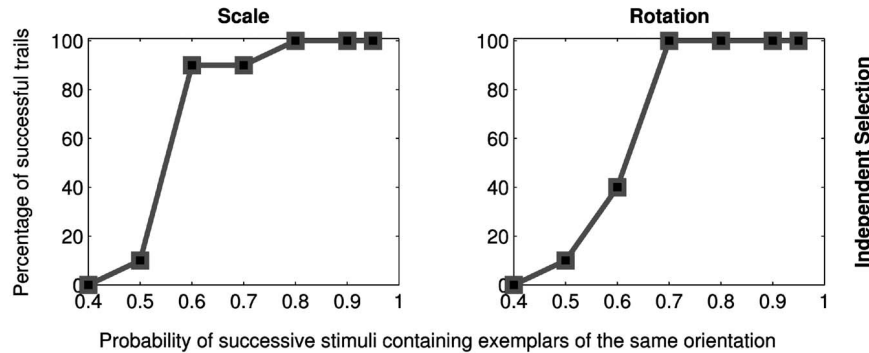


Fig. 6. (a) Reliability of learning representations of letters invariant to scale and (b) rotation, using the proposed algorithm. Each graph shows the percentage of trials for which the hierarchy successfully learned distinct representations for different letters. Within each graph, the number of successful trials is plotted for varying the probability that successive images will contain the same letter.

letter, the hierarchy was trained 10 times using different randomly generated sequences of input patterns. Training occurred for 2,500 iterations in each case. The number of these trials for which the task was successfully learned is shown in Fig. 6. It can be seen that the results are similar to those obtained above for learning different bar orientations invariant to translation.

4 CONCLUSIONS

Generalization is a vital component of intelligence. In vision, generalization underlies the formation of concepts by enabling the categorization of perceptually distinct objects. In addition, generalization underlies the recognition of individual objects by enabling identification despite changes in appearance. While much fruitful work in machine vision has explored the use of perspective invariants as a mechanism for object recognition, this does not appear to be the method employed by the brain [27]. In the cortex, object representations are built up over a hierarchy of processing stages through learning [13], [20], [23], [52] and the viewpoint invariance of these representations results from learning associations across time [22], [34], [37], [41], [50], [51], [53]. Many neural network models have been proposed which exploit this form of learning to develop invariant object representations. However, existing algorithms require training to be performed with isolated stimuli presented against blank backgrounds. This paper has suggested a simple modification to such methods that enables learning to succeed when the training environment contains multiple, co-occurring, stimuli. This suggested modification biases learning so that coincident objects will be represented by distinct nodes. The proposed algorithm also improves on previous methods by employing a more appropriate combination function and by learning correlations between consecutive image pairs rather than across sequences of images. It was shown that these modifications improved performance on a simple task where training data could contain co-occurring stimuli such that the reliability with which learning succeeded went from zero to 100 percent, across a range of conditions. Performance was also improved in experiments where stimuli were presented in isolation. The proposed algorithm thus provides a more robust method of learning invariant representations and could form the basis for the development of more powerful algorithms for learning invariance in real-world applications.

ACKNOWLEDGMENTS

This work was funded by EPSRC Grant number GR/S81339/01.

REFERENCES

- [1] H.B. Barlow, "Conditions for Versatile Learning, Helmholtz's Unconscious Inference, and the Task of Perception," *Vision Research*, vol. 30, pp. 1561-1571, 1990.
- [2] M.S. Bartlett and T.J. Sejnowski, "Unsupervised Learning of Invariant Representations of Faces through Temporal Association," *Computational Neuroscience: Int'l Rev. Neurobiology*, J. Bower, ed., vol. supplement 1, pp. 317-22, 1996.
- [3] M.S. Bartlett and T.J. Sejnowski, "Learning Viewpoint Invariant Face Representations from Visual Experience in an Attractor Network," *Network: Computation in Neural Systems*, vol. 9, no. 3, pp. 1-19, 1998.
- [4] S. Becker, "Learning to Categorize Objects Using Temporal Coherence," *Advances in Neural Information Processing Systems 5*, S.J. Hanson, J.D. Cowan, and C.L. Giles, eds. pp. 361-368, San Francisco: Morgan Kaufmann, 1993.
- [5] S. Becker, "Implicit Learning in 3D Object Recognition: The Importance of Temporal Context," *Neural Computation*, vol. 11, no. 2, pp. 347-374, 1999.
- [6] M.C.A. Booth and E.T. Rolls, "View-Invariant Representations of Familiar Objects by Neurons in the Inferior Temporal Visual Cortex," *Cerebral Cortex*, vol. 8, pp. 510-523, 1998.
- [7] A. Clark and C. Thornton, "Trading Spaces: Computation, Representation and the Limits of Uninformed Learning," *Behavioural and Brain Sciences*, vol. 20, no. 1, pp. 57-66, 1997.
- [8] M. Ebdon, "Towards a General Theory of Cerebral Neocortex," PhD thesis, Univ. of Sussex, U.K., 1996.
- [9] P. Földiák, "Forming Sparse Representations by Local Anti-Hebbian Learning," *Biological Cybernetics*, vol. 64, pp. 165-170, 1990.
- [10] P. Földiák, "Learning Invariance from Transformation Sequences," *Neural Computation*, vol. 3, pp. 194-200, 1991.
- [11] K. Fukushima, "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position," *Biological Cybernetics*, vol. 36, vol. 4, pp. 193-202, 1980.
- [12] K. Fukushima, "Neocognitron: A Hierarchical Neural Network Capable of Visual Pattern Recognition," *Neural Networks*, vol. 1, no. 2, pp. 119-130, 1988.
- [13] C.D. Gilbert, "Plasticity in Visual Perception and Physiology," *Current Opinion in Neurobiology*, vol. 6, no. 2, pp. 269-274, 1996.
- [14] M.A. Goodale and A.D. Milner, "Separate Visual Pathways for Perception and Action," *Trends in Neurosciences*, vol. 15, pp. 20-25, 1992.
- [15] G.E. Hinton, "Connectionist Learning Procedures," *Artificial Intelligence*, vol. 40, nos. 1-3, pp. 185-234, 1989.
- [16] D.H. Hubel and T.N. Wiesel, "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex," *J. Physiology (London)*, vol. 160, pp. 106-154, 1962.

- [17] D.H. Hubel and T.N. Wiesel, "Functional Architecture of Macaque Monkey Visual Cortex," *Proc. Royal Soc. London, Series B*, vol. 198, pp. 1-59, 1977.
- [18] E. Kobatake and K. Tanaka, "Neuronal Selectivities to Complex Object Features in the Ventral Visual Pathway of the Macaque Cerebral Cortex," *J. Neurophysiology*, vol. 71, no. 3, pp. 856-867, 1994.
- [19] K.P. Körding and P. König, "Neurons with Two Sites of Synaptic Integration Learn Invariant Representations," *Neural Computation*, vol. 13, no. 12, pp. 2823-2849, 2001.
- [20] N. Logothetis, "Object Vision and Visual Awareness," *Current Opinion in Neurobiology*, vol. 8, no. 4, pp. 536-544, 1998.
- [21] N. Logothetis and D.L. Sheinberg, "Visual Object Recognition," *Ann. Rev. Neuroscience*, vol. 19, pp. 577-621, 1996.
- [22] Y. Miyashita, "Neural Correlate of Visual Associative Long-Term Memory in the Primate Temporal Cortex," *Nature*, vol. 335, pp. 817-820, 1988.
- [23] V.B. Mountcastle, *Perceptual Neuroscience: The Cerebral Cortex*. Cambridge, Mass.: Harvard Univ. Press, 1998.
- [24] M.W. Oram and P. Földiák, "Learning Generalisation and Localisation: Competition for Stimulus Type and Receptive Field," *Neurocomputing*, vol. 11, nos. 2-4, pp. 297-321, 1996.
- [25] R.C. O'Reilly and M.H. Johnson, "Object Recognition and Sensitive Periods: A Computational Analysis of Visual Imprinting," *Neural Computation* vol. 6, pp. 357-389, 1994.
- [26] R.C. O'Reilly and J.L. McClelland, "The Self-Organization of Spatially Invariant Representations," Technical Report PDP.CNS.92.5, Dept. of Psychology, Carnegie Mellon Univ., 1992.
- [27] S.E. Palmer, *Vision Science: Photons to Phenomenology*. Cambridge, Mass.: MIT Press, 1999.
- [28] D.I. Perrett, "View-Dependent Coding in the Ventral Stream and Its Consequences for Recognition," *Vision and Movement Mechanisms in the Cerebral Cortex*, R. Caminiti, K.-P. Hoffmann, F. Lacquaniti, and J. Altman, eds., pp. 142-151. Strasbourg: HFSP, 1996.
- [29] D.I. Perrett, J.K. Hietanen, M.W. Oram, and P.J. Benson, "Organisation and Functions of Cells Responsive to Faces in the Temporal Cortex," *Philosophical Trans. Royal Soc. London*, vol. 335, pp. 23-30, 1992.
- [30] M. Riesenhuber and T. Poggio, "Are Cortical Models Really Bound by the 'Binding Problem'?" *Neuron*, vol. 24, no. 1, pp. 87-93, 1999.
- [31] M. Riesenhuber and T. Poggio, "Hierarchical Models of Object Recognition in Cortex," *Nature Neuroscience*, vol. 11, pp. 1019-1025, 1999.
- [32] E.T. Rolls, "Functions of the Primate Temporal Lobe Cortical Visual Areas in Invariant Visual Object and Face Recognition," *Neuron*, vol. 27, pp. 205-218, 2000.
- [33] E.T. Rolls and T. Milward, "A Model of Invariant Object Recognition in the Visual System: Learning Rules, Activation Functions, Lateral Inhibition, and Information-Based Performance Measures," *Neural Computation*, vol. 12, no. 11, pp. 2547-2572, 2000.
- [34] P. Sinha and T. Poggio, "Role of Learning in Three-Dimensional Form Perception," *Nature*, vol. 384, pp. 460-463, 1996.
- [35] M.W. Spratling and M.H. Johnson, "Pre-Integration Lateral Inhibition Enhances Unsupervised Learning," *Neural Computation*, vol. 14, vol. 9, pp. 2157-2179, 2002.
- [36] M.W. Spratling and M.H. Johnson, "Neural Coding Strategies and Mechanisms of Competition," *Cognitive Systems Research*, vol. 5, no. 2, pp. 93-117, 2004.
- [37] J. Stone, "Object Recognition Using Spatio-Temporal Signatures," *Vision Research*, vol. 38, no. 7, pp. 947-51, 1998.
- [38] J. Stone and A. Bray, "A Learning Rule for Extracting Spatio-Temporal Invariances," *Network: Computation in Neural Systems*, vol. 6, no. 3, pp. 429-436, 1995.
- [39] J.V. Stone, "A Canonical Microfunction for Learning Perceptual Invariances," *Perception*, vol. 25, pp. 207-220, 1996.
- [40] S.M. Stringer and E.T. Rolls, "Position Invariant Recognition in the Visual System with Cluttered Environments," *Neural Networks*, vol. 13, pp. 305-315, 2000.
- [41] M.P. Stryker, "Temporal Associations," *Nature*, vol. 354, pp. 108-109, 1991.
- [42] K. Tanaka, "Representation of Visual Feature Objects in the Inferotemporal Cortex," *Neural Networks*, vol. 9, no. 8, pp. 1459-1475, 1996.
- [43] J.N. Templeman and M.H. Loew, "Staged Assimilation: A System for Detecting Invariant Features in Temporally Coherent Visual Stimuli," *Proc. Int'l Joint Conf. Neural Networks*, vol. 1, pp. 731-738, 1989.
- [44] C. Thornton, "Re-Presenting Representation," *Forms of Representation: An Interdisciplinary Theme for Cognitive Science*, D.M. Peterson, ed., pp. 152-162. Exeter, U.K.: Intellect Books, 1996.
- [45] M.J. Tovee, E.T. Rolls, and P. Azzopardi, "Translation Invariance in the Responses to Faces of Single Neurons in the Temporal Visual Cortical Areas of the Alert Macaque," *J. Neurophysiology*, vol. 72, no. 3, pp. 1049-1060, 1994.
- [46] L.G. Ungerleider and M. Mishkin, "Two Cortical Visual Systems," *Analysis of Visual Behavior*, D.J. Ingle, M.A. Goodale, and R.J.W. Mansfield, eds., pp. 549-586. Cambridge, Mass.: MIT Press, 1982.
- [47] G. Wallis, "Neural Mechanisms Underlying Processing in the Visual Areas of the Occipital and Temporal Lobes," PhD thesis, Corpus Christi College/Dept. of Experimental Psychology, Univ. of Oxford, U.K., 1994.
- [48] G. Wallis, "Using Spatio-Temporal Correlations to Learn Invariant Object Recognition," *Neural Networks*, vol. 9, no. 9, pp. 1513-1519, 1996.
- [49] G. Wallis, "Spatio-Temporal Influences at the Neural Level of Object Recognition," *Network: Computation in Neural Systems*, vol. 9, no. 2, pp. 265-278, 1998.
- [50] G. Wallis, "Temporal Order in Human Object Recognition," *J. Biological Systems*, vol. 6, no. 3, pp. 299-313, 1998.
- [51] G. Wallis, "The Role of Object Motion in Forging Long-Term Representations of Objects," *Visual Cognition*, vol. 9, pp. 233-247, 2002.
- [52] G. Wallis and H. Bülthoff, "Learning to Recognize Objects," *Trends in Cognitive Sciences*, vol. 3, no. 1, pp. 22-31, 1999.
- [53] G. Wallis and H. Bülthoff, "Role of Temporal Association in Establishing Recognition Memory," *Proc. Nat'l Academy of Sciences USA*, vol. 98, no. 8, pp. 4800-4804, 2001.
- [54] G. Wallis, E. Rolls, and P. Földiák, "Learning Invariant Responses to the Natural Transformations of Objects," *Proc. Int'l Joint Conf. Neural Networks*, vol. 2, pp. 1087-1090, 1993.
- [55] G. Wallis and E.T. Rolls, "Invariant Face and Object Recognition in the Visual System," *Progress in Neurobiology*, vol. 51, no. 2, pp. 167-194, 1997.
- [56] L. Wiskott and T.J. Sejnowski, "Slow Feature Analysis: Unsupervised Learning of Invariances," *Neural Computation*, vol. 14, no. 4, pp. 715-770, 2002.



Michael W. Spratling received the BEng degree in engineering science from Loughborough University and the MSc and PhD degrees in artificial intelligence from the University of Edinburgh. He has held several research positions, both in psychology and in engineering, and he is currently a lecturer in the Division of Engineering at King's College London and a research fellow in the Centre for Brain and Cognitive Development at Birkbeck College, London.

► For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.