IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(5):753–761, 2005

Learning Viewpoint Invariant Perceptual Representations from Cluttered Images

M. W. Spratling
Division of Engineering, King's College, London. UK.
and Centre for Brain and Cognitive Development, Birkbeck College, London. UK.

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 generalise 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 are presented in isolation and is therefore unlikely to succeed in real-world applications where multiple objects can co-occur in the visual input. This article 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 primary visual cortex to the temporal lobe (Ungerleider and Mishkin, 1982; Goodale and Milner, 1992). 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 (Hubel and Wiesel, 1977), while neurons in temporal areas are selective to complex objects, such as faces, appearing anywhere in the visual field (Perrett et al., 1992; Tovee et al., 1994; Tanaka, 1996; Perrett, 1996; Logothetis and Sheinberg, 1996; Booth and Rolls, 1998; Rolls, 2000). Hence, cortical cells along the ventral pathway learn increasing specificity together with increasing invariance (Kobatake and Tanaka, 1994; Tovee et al., 1994). 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 learnt from lower-level ones, and this process can be repeated hierarchically, such that at each stage neurons become tuned to ever more specialised and invariant features of the environment (Clark and Thornton, 1997; Thornton, 1996). This process makes tractable the task of learning complex perceptual representations (Clark and Thornton, 1997) 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 co-active lower-level features. A node must thus learn to represent a *conjunction* of pre-synaptic inputs. In contrast, a more invariant representation results from a node responding to multiple, non-coactive, lower-level features. A node must thus learn to represent a *disjunction* of pre-synaptic inputs. Hence, several existing architectures for invariant object recognition (*e.g.*, the Neocognitron (Fukushima, 1980, 1988), and the HMAX model (Riesenhuber and Poggio, 1999a,b)) consist of alternating stages of neural populations that perform these two operations. A simple, two stage, neural hierarchy of this kind is shown in figure 1. It has been proposed (Fukushima, 1988; Templeman and Loew, 1989; Riesenhuber and Poggio, 1999b) that these two forms of processing correspond to the functionality of simple and complex cells observed in the primary visual cortex (Hubel and Wiesel, 1962).

What operation is performed by a node depends critically 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 pre-synaptic 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 (Riesenhuber and Poggio, 1999a,b). 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 learnt at different locations and fixed weights are used to pool responses from these nodes to achieve translation invariance (Fukushima, 1980, 1988). 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 (Riesenhuber and Poggio, 1999a,b). Hence, viewpoint invariance is built in to both of these architectures, rather than being learnt.

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 (Földiák, 1990; Barlow, 1990). To learn a disjunction, it is necessary for a node to form strong connection weights with a set of non-coactive inputs. The 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” (Wallis, 2002). By doing so, a neuron learns to become selective for statistical regularities across time. This suggestion (Hinton, 1989) has formed the basis for a large number of algorithms that learn invariance from sequences of images (*e.g.*, Templeman and Loew, 1989; Földiák, 1991; O’Reilly and McClelland, 1992; Becker, 1993, 1999; O’Reilly and Johnson, 1994; Stone and Bray, 1995; Stone, 1996; Ebdon, 1996; Oram and Földiák, 1996; Wallis et al., 1993; Wallis, 1994, 1996, 1998a; Wallis and Rolls, 1997; Bartlett and Sejnowski, 1996, 1998; Stringer and Rolls, 2000; Rolls and Milward, 2000; Körding and König, 2001; Wiskott and Sejnowski, 2002). 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 learnt from temporal associations (Wallis, 1998b, 2002; Wallis and Bülthoff, 2001; Sinha and Poggio, 1996; Stone, 1998; Stryker, 1991; Miyashita, 1988). 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 for instantaneous activity. This enables previous activity to influence learning of subsequent input patterns and hence transforms temporal regularities into spatial ones (Wallis, 1994).

Despite the large number 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 (Földiák, 1991; Oram and Földiák, 1996; Wallis, 1996; Stringer and Rolls, 2000), 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 can not 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 equation 6 below). 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 be-

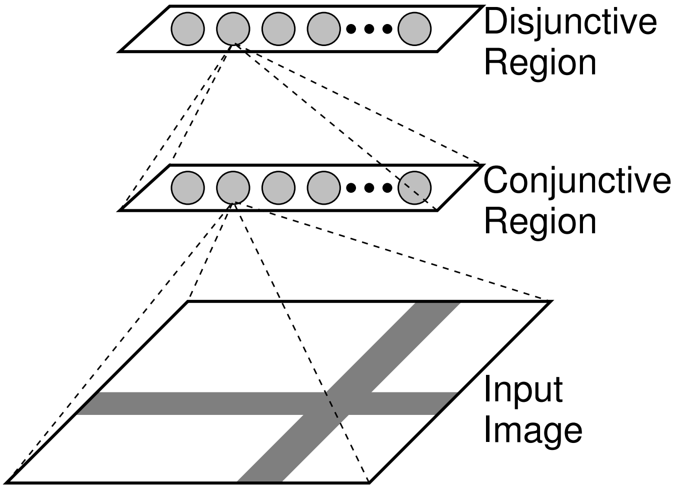


Figure 1: The proposed neural network architecture for learning perceptual invariances. Nodes in the lower region receive connections from an input image, and learn to represent conjunctions of these pre-synaptic inputs, thus forming more specific representations. Nodes in the upper region receive connections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

tween the current input and the output generated in response to the previous image. Hence, temporal associations are learnt 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 (2000), 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 affects 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 (Oram and Földiák, 1996; Wallis and Rolls, 1997; Rolls and Milward, 2000).

Several previous models which learn invariance have calculated node activity using a weighted sum of the inputs (*e.g.*, Földiák, 1991; Oram and Földiák, 1996). As mentioned in the introduction the $\max$ operator has been proposed as a more appropriate combination function for generating invariant responses (Riesenhuber and Poggio, 1999a,b). Summation allows multiple, non-optimal, 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 equation 1 below). 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.*, Riesenhuber and Poggio, 1999a,b), while those models that do learn invariances use an inappropriate combination function (*e.g.*, Földiák, 1991; Oram and Földiák, 1996).

2.1 Implementation Details

The performance of the proposed algorithm was explored using a simple neural hierarchy consisting of two stages, or regions (see figure 1). One region learnt conjunctions and the other region learnt disjunctions. The conjunctive region was implemented using the algorithm presented in Spratling and Johnson (2002, 2004), 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:

$$y_{jk}^{t} = \max_{i=1}^{m} \left\{ Z_{ijk}^{t} \right\} \tag{1}$$

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

$$y_{jk}^{t} = \sum_{i=1}^{m} \left(w_{ijk} X_{ijk}^{t} \right) \tag{2}$$

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 normalised 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 pre-integration lateral inhibition:

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

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

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 pre-synaptic activity values (x_{ijk}^{t}) are either the activations of nodes in the lower region at the previous time step (*i.e.*, $y_{j'k-1}^{t-1}$) or are external, sensory, inputs supplied to the hierarchy.

The value of Z_{ijk}^{t} depends upon the strength of pre-synaptic activity and the strength of the weight received from that input (i). This value is adjusted using both post- and pre-synaptic weight normalisation. Such normalisation 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 normalisation, biases the node to respond to an active input to which it has previously responded. While, the second form of normalisation, 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 pre-synaptic activity and the strength of the lateral inhibition directed towards 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 article, the value of α^{t} was increased from zero to ten 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:

$$y_{jk}^{t} = y_{jk}^{t} \left(1 + \rho \right) \tag{5}$$

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:

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

Where γ is a parameter controlling the learning rate ($\gamma = \frac{1}{4}$ was used in the simulations presented in this article), 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 non-coactive 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 normalised such that $\sum_{j=1}^{n} w_{ijk} = 1$. This normalisation 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:

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

Where β is a parameter controlling the learning rate ($\beta = 1$ was used in the simulations presented in this article), $\bar{x}_{jk}$ 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 normalised 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 post-synaptic 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:

$$\Delta w_{ijk} = \frac{\beta\left(X_{ijk} - 0.5x_{ijk}\right)^{-}}{\sum_{p=1}^{n} y_{pk}} \quad \left(y_{jk} - \bar{y}_{k}\right) \tag{8}$$

Where X_{ijk} is the input activation from source i to node j after inhibition (see equation 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 equation 7 (or prior application of equation 8). Negative weights are generated when a node is active and inputs, which 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, re-occurs 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 towards 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 equation 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 articles (*e.g.*, Földiák, 1991; Wallis, 1994, 1996)

To compare the results of the proposed algorithm with previous methods, certain experiments were also repeated, using the algorithm proposed in (Földiák, 1991) for the disjunctive region. Hence, equation 6 was replaced by:

$$\Delta w_{ijk} = 0.02\left(x_{ijk} - w_{ijk}\right) y^*_{jk}$$

where y^*_{jk} was a trace of previous activity, calculated as above, and equation 1 was replaced by:

$$y^t_{jk} = \sum_{i=1}^{m} \left(w_{ijk} x^t_{ijk}\right)$$

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.*, Wallis, 1994, 1996; Oram and Földiák, 1996; Stringer and Rolls, 2000), 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 (Földiák, 1991; Oram and Földiák, 1996; Körding and König, 2001). A two-region hierarchy of neural networks was used. The lower region received input from images of training stimuli and learnt conjunctions of features within these images, while the upper region received input from all the nodes in the lower region and learnt 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 region. Learning, in both networks proceeded from the start of training (*i.e.*, there was no pre-training of the lower network before training the upper network).

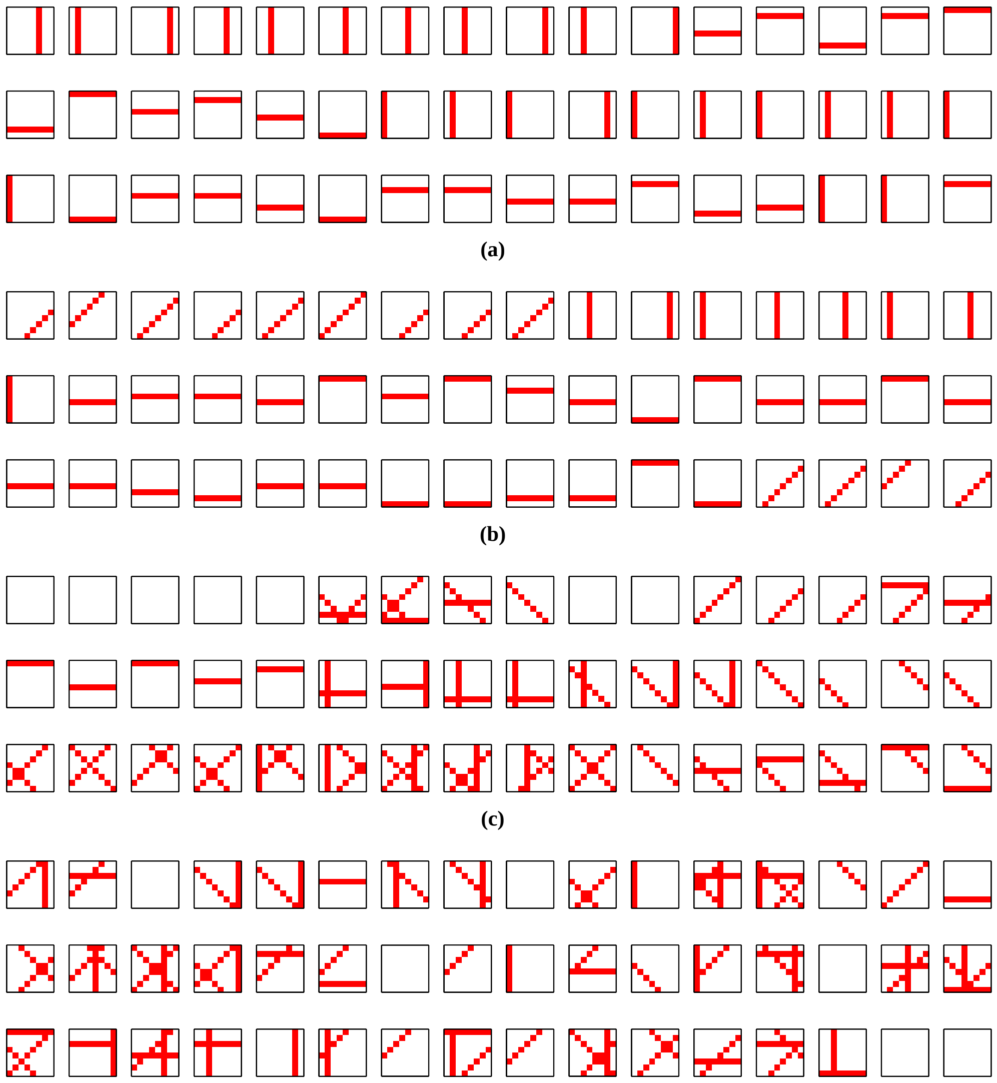
Training data consisted of a series eight-by-eight pixel images, within which one-pixel wide horizontal, vertical and diagonal bars could appear. Two methods for generating the training data were used. Bars could be selected such that orientations were mutually exclusive or the bar orientations present 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 bars[1] in others). Figure 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. Figure 3 shows a typical example of the weights learnt 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 ten times using different randomly generated sequences of input patterns. The number of these trials for which the task was successfully learnt, are shown in figure 4. For trails in which two orientations were present the hierarchy was trained for 5000 iterations, and for trials in which four orientations were used, since there were twice as many representations to be learnt, the hierarchy was trained for 10000 iterations.



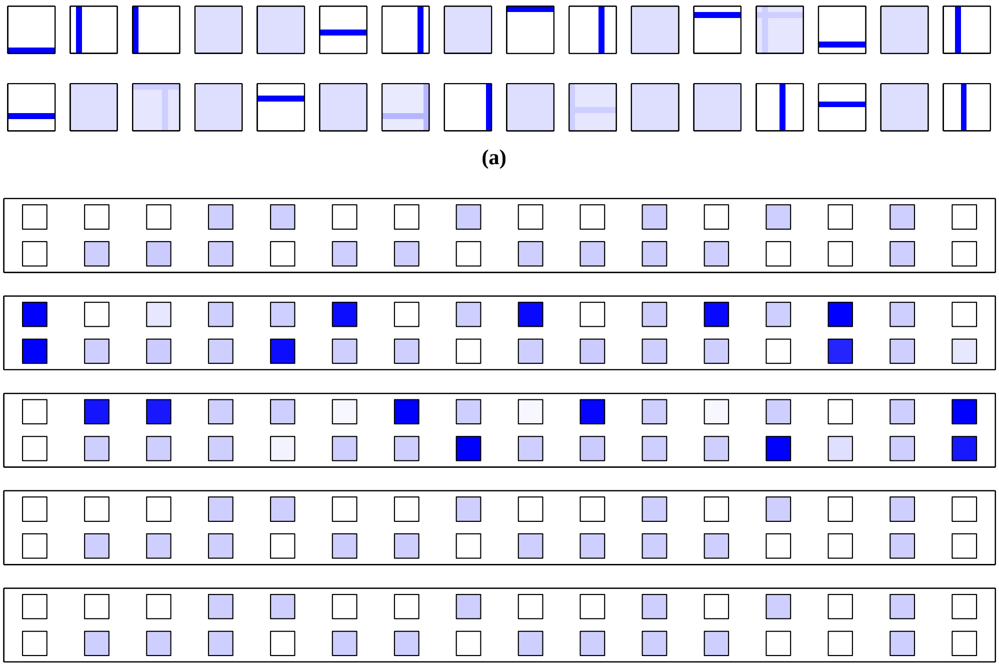
(a)

(b)

(c)

(d)

Figure 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).



(a)

(b)

Figure 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, learnt after 5000 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 as 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 learnt to represent horizontal bars, while another has learnt to represent vertical bars: each of these nodes is selective for all the nodes in the lower network which represent bars of that orientation.

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 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 figure 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 ten 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 when multiple object representations are active in response to a single image, that 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 ten 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 figure 4. It can be seen that when images may contain multiple views of the same object, a reduction in the

[1] the seven longest diagonals in each direction.

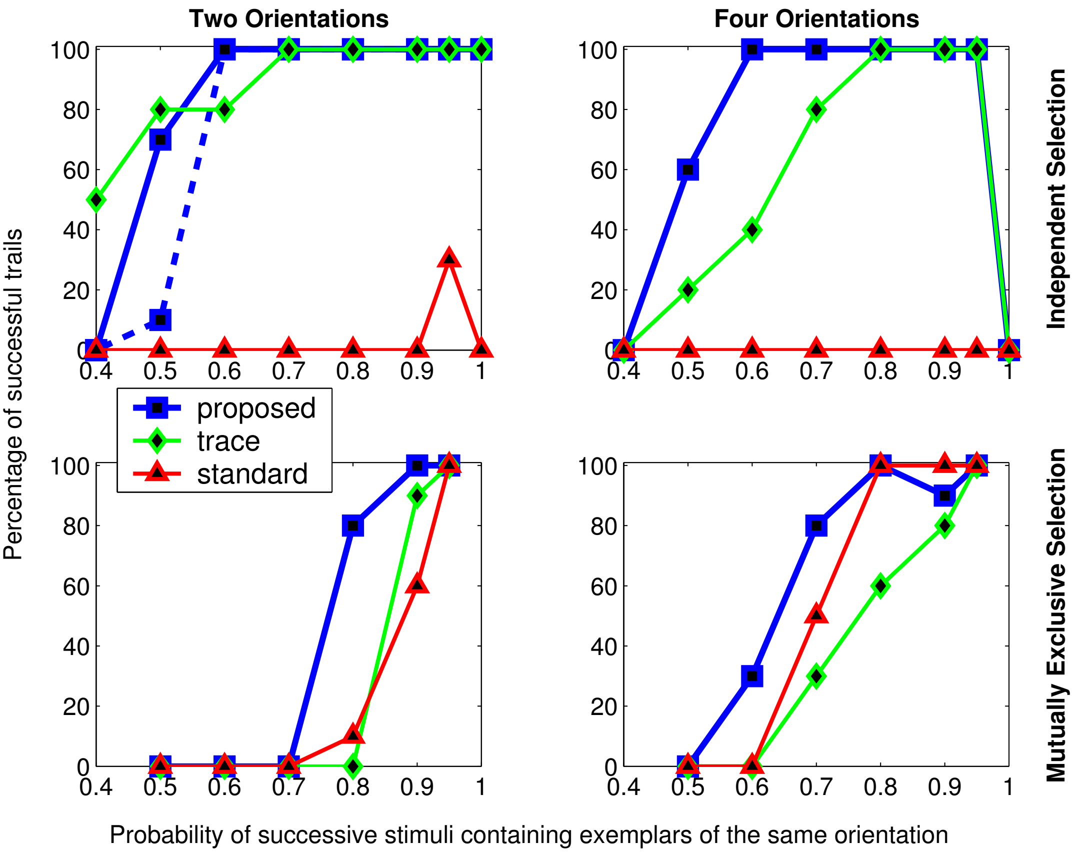


Figure 4: Reliability of learning representations of line orientation invariant to translation. Each graph shows the percentage of trails for which the hierarchy successfully learnt distinct representations for bars of each orientation, at all image locations. Within each graph, the number of successful 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% of images contained two bars of the same orientation.

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 equation 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% 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 were 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 figure 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 letter, the hierarchy was trained ten times using different randomly generated sequences of input patterns. Training occurred for 2500 iterations, in each case. The number of these trials for which the task was successfully learnt, are shown in figure 6. It can be seen that the results are similar to those obtained above for learning different bar orientations invariant to translation.

4 Conclusions

Generalisation is a vital component of intelligence. In vision, generalisation underlies the formation of concepts by enabling the categorisation of perceptually distinct objects. In addition, generalisation 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 (Palmer, 1999). In the cortex, object representations are built up over a hierarchy of processing stages through learning (Gilbert, 1996; Logothetis, 1998; Mountcastle, 1998; Wallis and Bülthoff, 1999) and the viewpoint invariance of these representations results from learning associations across time (Wallis, 1998b, 2002; Wallis and Bülthoff, 2001; Sinha and Poggio, 1996; Stone, 1998; Stryker, 1991; Miyashita, 1988). 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 article 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.

Acknowledgements

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

References

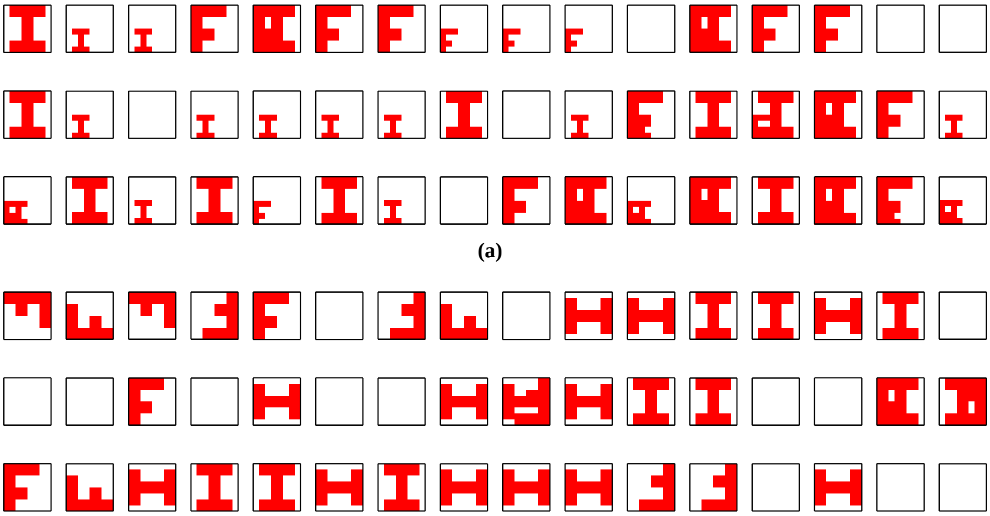
Barlow, H. B. (1990). Conditions for versatile learning, Helmholtz’s unconscious inference, and the task of perception. *Vision Research*, 30:1561–71.

Bartlett, M. S. and Sejnowski, T. J. (1996). Unsupervised learning of invariant representations of faces through temporal association. In Bower, J., editor, *Computational Neuroscience: International Review of Neurobiology*, volume Suppliment 1, pages 317–22, San Diego, CA. Academic Press.

Bartlett, M. S. and Sejnowski, T. J. (1998). Learning viewpoint invariant face representations from visual experience in an attractor network. *Network: Computation in Neural Systems*, 9(3):1–19.

Becker, S. (1993). Learning to categorize objects using temporal coherence. In Hanson, S. J., Cowan, J. D., and Giles, C. L., editors, *Advances in Neural Information Processing Systems 5*, pages 361–8, San Francisco, CA. Morgan Kaufmann.

Becker, S. (1999). Implicit learning in 3D object recognition: the importance of temporal context. *Neural Computation*, 11(2):347–74.



(a)

(b)

Figure 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.

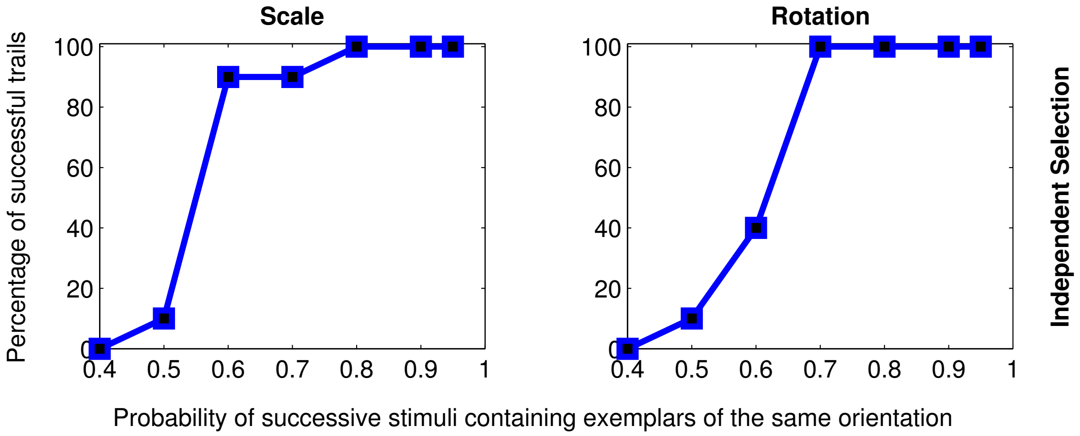


Figure 6: Reliability of learning representations of letters invariant to scale (left) and rotation (right), using the proposed algorithm. Each graph shows the percentage of trails for which the hierarchy successfully learnt distinct representations for different letters. Within each graph, the number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

Booth, M. C. A. and Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex*, 8:510–23.

Clark, A. and Thornton, C. (1997). Trading spaces: computation, representation and the limits of uninformed learning. *Behavioural and Brain Sciences*, 20(1):57–66.

Ebdon, M. (1996). *Towards a General Theory of Cerebral Neocortex*. PhD thesis, University of Sussex, UK.

Földiák, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*, 64:165–70.

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(4):193–202.

Fukushima, K. (1988). Neocognitron: a hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1(2):119–30.

Gilbert, C. D. (1996). Plasticity in visual perception and physiology. *Current Opinion in Neurobiology*, 6(2):269–74.

Goodale, M. A. and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15:20–5.

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence*, 40(1-3):185–234.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology (London)*, 160:106–54.

Hubel, D. H. and Wiesel, T. N. (1977). Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London. Series B*, 198:1–59.

Kobatake, E. and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, 71(3):856–67.

Körding, K. P. and König, P. (2001). Neurons with two sites of synaptic integration learn invariant representations. *Neural Computation*, 13(12):2823–49.

Logothetis, N. (1998). Object vision and visual awareness. *Current Opinion in Neurobiology*, 8(4):536–44.

Logothetis, N. and Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, 19:577–621.

Miyashita, Y. (1988). Neural correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, 335:817–20.

Mountcastle, V. B. (1998). *Perceptual Neuroscience: The Cerebral Cortex*. Harvard University Press, Cambridge, MA.

Oram, M. W. and Földiák, P. (1996). Learning generalisation and localisation: competition for stimulus type and receptive field. *Neurocomputing*, 11(2-4):297–321.

O'Reilly, R. C. and Johnson, M. H. (1994). Object recognition and sensitive periods: a computational analysis of visual imprinting. *Neural Computation*, 6:357–89.

O'Reilly, R. C. and McClelland, J. L. (1992). The self-organization of spatially invariant representations. Technical Report PDP.CNS.92.5, Department of Psychology, Carnegie Mellon University.

Palmer, S. E. (1999). *Vision Science: Photons to Phenomenology*. MIT Press, Cambridge, MA.

Perrett, D. I. (1996). View-dependent coding in the ventral stream and its consequences for recognition. In Caminiti, R., Hoffmann, K.-P., Lacquaniti, F., and Altman, J., editors, *Vision and Movement Mechanisms in the Cerebral Cortex*, pages 142–51. HFSP, Strasbourg.

Perrett, D. I., Hietanen, J. K., Oram, M. W., and Benson, P. J. (1992). Organisation and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London*, 335:23–30.

Riesenhuber, M. and Poggio, T. (1999a). Are cortical models really bound by the "binding problem"? *Neuron*, 24(1):87–93.

Riesenhuber, M. and Poggio, T. (1999b). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–25.

Rolls, E. T. (2000). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron*, 27:205–18.

Rolls, E. T. and Milward, T. (2000). A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Computation*, 12(11):2547–72.

Sinha, P. and Poggio, T. (1996). Role of learning in three-dimensional form perception. *Nature*, 384:460–3.

Spratling, M. W. and Johnson, M. H. (2002). Pre-integration lateral inhibition enhances unsupervised learning. *Neural Computation*, 14(9):2157–79.

Spratling, M. W. and Johnson, M. H. (2004). Neural coding strategies and mechanisms of competition. *Cognitive Systems Research*, 5(2):93–117.

Stone, J. (1998). Object recognition using spatio- temporal signatures. *Vision Research*, 38(7):947–51.

Stone, J. and Bray, A. (1995). A learning rule for extracting spatio-temporal invariances. *Network: Computation in Neural Systems*, 6(3):429–36.

Stone, J. V. (1996). A canonical microfunction for learning perceptual invariances. *Perception*, 25:207–20.

Stringer, S. M. and Rolls, E. T. (2000). Position invariant recognition in the visual system with cluttered environments. *Neural Networks*, 13:305–15.

Stryker, M. P. (1991). Temporal associations. *Nature*, 354:108–9.

Tanaka, K. (1996). Representation of visual feature objects in the inferotemporal cortex. *Neural Networks*, 9(8):1459–75.

Templeman, J. N. and Loew, M. H. (1989). Staged assimilation: a system for detecting invariant features in temporally coherent visual stimuli. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN89)*, volume 1, pages 731–8, New York, NY. IEEE Press.

Thornton, C. (1996). Re-presenting representation. In Peterson, D. M., editor, *Forms of Representation: An Interdisciplinary Theme for Cognitive Science*, pages 152–62. Intellect Books, Exeter, UK.

Tovee, M. J., Rolls, E. T., and Azzopardi, P. (1994). Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque. *Journal of Neurophysiology*, 72(3):1049–60.

Ungerleider, L. G. and Mishkin, M. (1982). Two cortical visual systems. In Ingle, D. J., Goodale, M. A., and Mansfield, R. J. W., editors, *Analysis of Visual Behavior*, pages 549–86. MIT Press, Cambridge, MA.

Wallis, G. (1994). *Neural Mechanisms Underlying Processing in the Visual Areas of the Occipital and Temporal Lobes*. PhD thesis, Corpus Christi College/Department of Experimental Psychology, University of Oxford, UK.

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

Wallis, G. (1998a). Spatio-temporal influences at the neural level of object recognition. *Network: Computation in Neural Systems*, 9(2):265–78.

Wallis, G. (1998b). Temporal order in human object recognition. *Journal of Biological Systems*, 6(3):299–313.

Wallis, G. (2002). The role of object motion in forging long-term representations of objects. *Visual Cognition*, 9:233–47.

Wallis, G. and Bülthoff, H. (1999). Learning to recognize objects. *Trends in Cognitive Sciences*, 3(1):22–31.

Wallis, G. and Bülthoff, H. (2001). Role of temporal association in establishing recognition memory. *Proceedings of the National Academy of Sciences USA*, 98(8):4800–4.

Wallis, G., Rolls, E., and Földiák, P. (1993). Learning invariant responses to the natural transformations of objects. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2, pages 1087–90.

Wallis, G. and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, 51(2):167–94.

Wiskott, L. and Sejnowski, T. J. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Computation*, 14(4):715–70.