Available online at www.sciencedirect.com
ScienceDirect
Vision Research 46 (2006) 3994–4006
Vision Research
www.elsevier.com/locate/visres

Spatial vs temporal continuity in view invariant visual object recognition learning

Gavin Perry, Edmund T. Rolls *, Simon M. Stringer

Oxford University, Centre for Computational Neuroscience, Department of Experimental Psychology, South Parks Road, Oxford OX1 3UD, UK

Received 8 February 2006; received in revised form 22 June 2006

Abstract

We show in a 4-layer competitive neuronal network that continuous transformation learning, which uses spatial correlations and a purely associative (Hebbian) synaptic modification rule, can build view invariant representations of complex 3D objects. This occurs even when views of the different objects are interleaved, a condition where temporal trace learning fails. Human psychophysical experiments showed that view invariant object learning can occur when spatial but not temporal continuity applies because of interleaving of stimuli, although sequential presentation, which produces temporal continuity, can facilitate learning. Thus continuous transformation learning is an important principle that may contribute to view invariant object recognition.

Keywords: Object recognition; Continuous transformation; Trace learning; Inferior temporal cortex

1. Introduction

There is now much evidence demonstrating that over successive stages the ventral visual system develops neurons that respond to objects or faces with view, size, and position (translation) invariance (Rolls, 1992, 2000, 2006; Rolls & Deco, 2002; Desimone, 1991; Tanaka, Saito, Fukada, & Moriya, 1991). For example, it has been shown that the macaque inferior temporal visual cortex has neurons that respond to faces and objects with invariance to translation (Tovee, Rolls, & Azzopardi, 1994; Kobotake & Tanaka, 1994; Ito, Tamura, Fujita, & Tanaka, 1995; Op de Beeck & Vogels, 2000; Rolls, Aggelopoulos, & Zheng, 2003), size (Rolls & Baylis, 1986; Ito et al., 1995), contrast (Rolls & Baylis, 1986), lighting (Vogels & Biederman, 2002), spatial frequency (Rolls, Baylis, & Leonard, 1985; Rolls, Baylis, & Hasselmo, 1987), and view (Hasselmo, Rolls, Baylis, & Nalwa, 1989; Booth & Rolls, 1998). It is crucially important that the visual system builds invariant representations, for only then can one-trial learning about an object generalize usefully to other transforms of the same object (Rolls & Deco, 2002). Building invariant representations of objects is a major computational issue, and the means by which the cerebral cortex solves this problem is a topic of great interest (Riesenhuber & Poggio, 1999; Biederman, 1987; Ullman, 1996; Rolls & Deco, 2002).

One proposed method for the learning of invariance in the visual system is to utilize the temporal continuity of objects in the visual environment (over short time periods) to help the learning of invariant representations (Földiák, 1991; Rolls, 1992; Wallis & Rolls, 1997; Rolls & Milward, 2000; Rolls & Stringer, 2001). Temporal continuity can be utilized by, for example, associative learning rules that incorporate a temporal trace of activity in the post-synaptic neuron (Földiák, 1991; Rolls, 1992; Wallis & Rolls, 1997). These rules encourage neurons to respond to input patterns that occur close together in time, which, given the natural statistics of the visual world, are likely to represent different transforms (views) of the same object. Temporal continuity is also a feature of other proposals (Stone, 1996; Bartlett

* Corresponding author. Fax: +44 1865 310447.
E-mail address: Edmund.Rolls@psy.ox.ac.uk (E.T. Rolls).
URL: www.cns.ox.ac.uk (E.T. Rolls).

0042-6989/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2006.07.025

& Sejnowski, 1998; Becker, 1999; Einhäuser, Kayser, König, & Körding, 2002; Wiskott & Sejnowski, 2002).

Recently, spatial continuity in the different views of a transforming object has been proposed as another principle of invariance learning (Stringer, Perry, Rolls, & Proske, 2006). In continuous transformation (CT) learning a competitive network using an associative synaptic modification rule learns to respond to an initial view of an object, and then similar views activate the same post-synaptic neuron through the strengthened synapses. As the object transforms continuously, the different views become associated onto the same post-synaptic neurons, as illustrated in Fig. 2. The CT learning effect can operate even when there are large separations of time between the presentation of views of the same object, and even if views of different stimuli are presented during this intervening time period in an interleaved training condition (Stringer et al., 2006). Spatial continuity in the context of continuous transformation learning is the property that the different views of an object are sufficiently similar that after one view has been learned, an adjacent view will have sufficient overlap of the active inputs to activate the same neuron, as illustrated in Fig. 2. In topologically mapped systems, these adjacent (overlapping) inputs will be spatially close, but need not be in a non-topologically mapped system.

In this paper, we compare computer simulations with psychophysical studies using the same set of stimuli to investigate the relative contributions of temporal continuity and spatial continuity in the learning of view invariant representations of objects in the brain.

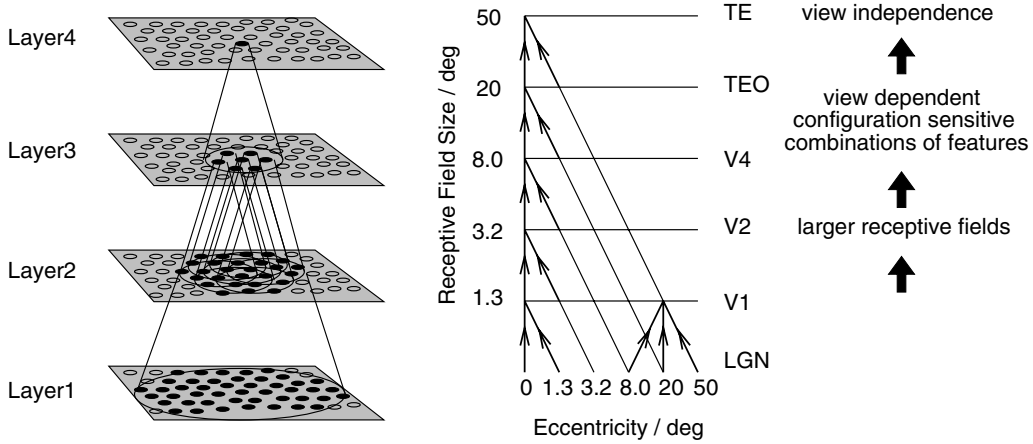
First, we test how closely predictions of the temporal vs spatial continuity theories are met in a hierarchical model of the ventral visual stream, VisNet (Wallis & Rolls, 1997; Rolls & Milward, 2000) illustrated in Fig. 1, in which the parameters can be precisely controlled. Use of the model helps to show the type of result expected if the system is trained with temporal trace vs spatial continuous transformation paradigms.

We then use the same realistically difficult set of objects in a psychophysical experiment with humans to investigate whether humans' learning reflects the use of short-term temporal correlations vs spatial continuity in the different transforms of each object.

2. Methods

In order to investigate the roles of temporal vs spatial continuity in human invariance learning in the context of the temporal trace and continuous transformation theories of invariance learning, human performance and a network model trained with the same set of stimuli, were compared with a range of training stimulus presentation paradigms. Key predictions of the continuous transformation (CT) vs temporal trace theories tested are that CT but not temporal trace learning can self-organize invariant representations when the views of different objects are interleaved, and that CT learning but not necessarily temporal trace learning will perform poorly if the spacing between the closest views become larger, thus breaking the spatial continuity in the images seen by the network.

In the 'interleaved' training condition, an initial view of the first object was shown, followed by an initial view of the second object and then an initial view of each of the remaining objects in order. Once a view of each object had been shown the next image in clockwise (viewed from above) sequence of the first object was presented followed by the next image in sequence of the second object, then the third and so on. This procedure ensured that in the interleaved condition two views of the same object did not occur close together in time. It is a prediction of the temporal trace hypothesis (and any model that uses temporal continuity to learn invariance) that training in this manner should cause invariance learning to be impaired (as views of different objects could become associated together



	Dimensions	# Connections	Radius
Layer 1	32x32	100	12
Layer 2	32x32	100	9
Layer 3	32x32	100	6
Layer 4	32x32	272	6
Retina	128x128x32	-	-

Fig. 1. (Left) Schematic diagram of the four layer hierarchical competitive network, VisNet. Convergence through the network is designed to provide fourth layer neurons with information from across the entire input retina. (Right) Convergence in the visual system. V1, visual cortex area V1; TEO, posterior inferior temporal cortex; TE, inferior temporal cortex (IT). (Bottom) Network dimensions showing the number of connections per neuron and the radius in the preceding layer from which 67% are received.

due to their close proximity). In contrast CT learning does not involve temporal associations, and the continuous transformation hypothesis predicts that learning should be unimpaired by interleaved training.

Training in a 'sequential' condition, in which during training all the views of one object are presented in the order in which they occur during transformation (e.g. rotation) of the object in the real world, before the views of another object are presented, was predicted to produce good learning by both the trace and the CT theories.

In a 'permuted' training paradigm used in some experiments, all the views of one object are presented before the views of another object are presented, but the different views within an object are presented in a permuted sequence. It is predicted by the temporal trace theory that learning could be better, because distant views of the same object can become temporally associated even if the trace does not last for longer than a few views during training. It is predicted that CT learning may be slower in the permuted condition, because successively presented views at least early in training may not be sufficiently spatially overlapping to activate the same output neurons (Stringer et al., 2006) (see Fig. 2).

It is also predicted that (at least in a previously untrained network) CT learning will be impaired if the different views of an object are too widely separated to overlap sufficiently to activate the same output neurons, whereas temporal trace learning should be less affected, as temporal associations can still be made between different views.

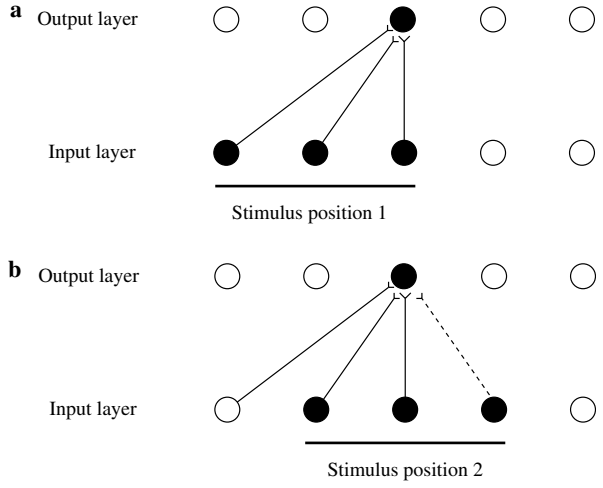


Fig. 2. An illustration of how CT learning would function in a network with a single layer of forward synaptic connections between an input layer of neurons and an output layer. Initially the forward synaptic weights are set to random values. (a) The initial presentation of a stimulus to the network in position 1. Activation from the (shaded) active input cells is transmitted through the initially random forward connections to stimulate the cells in the output layer. The shaded cell in the output layer wins the competition in that layer. The weights from the active input cells to the active output neuron are then strengthened using an associative learning rule. (b) What happens after the stimulus is shifted by a small amount to a new partially overlapping position 2. As some of the active input cells are the same as those that were active when the stimulus was presented in position 1, the same output cell is driven by these previously strengthened afferents to win the competition again. The rightmost shaded input cell activated by the stimulus in position 2, which was inactive when the stimulus was in position 1, now has its connection to the active output cell strengthened (denoted by the dashed line). Thus, the same neuron in the output layer has learned to respond to the two input patterns that have similar vector elements in common. As can be seen, the process can be continued for subsequent shifts, provided that a sufficient proportion of input cells stay active between individual shifts.

2.1. Stimuli

Stimuli that had a small rotation between different views as well as precisely controlled lighting and surface texture were created using the 3D Studio Max (Autodesk, Inc., San Rafael, CA) software package. Fig. 3 shows the five stimulus objects from a selection of views at 72° intervals. Each was created by combining a small number (either five or six, depending on the object) of simple 3D geometric shapes (e.g. cubes, cylinders, and wedges). Objects composed from simple geometric shapes such as these have been used in previous tests of viewpoint invariance (Biederman & Gerhardstein, 1993; Biederman & Bar, 1999). The images were produced in 256 level greyscale. All objects were rendered as if lit by a single light positioned directly behind the vantage point of the viewer, with no specularities or cast shadows. The surface reflectances of each adjacent constituent part within an object were made different so as to better highlight the edges between the parts and hence make the overall structure of the object clearer. A virtual camera was set up to circle each object around its vertical axis in one hundred steps, and a new image was rendered at each step, producing 100 images spanning 360° with 3.6° between adjacent views.

2.2. Modeling

The predictions and the training properties expected with the temporal trace and CT theories were tested by simulation in a model of the ventral visual system. This model, VisNet, was implemented following the proposals of Rolls (1992) by Wallis and Rolls (1997) and developed further by Rolls and Milward (2000), based on the following: (i) a series of hierarchical competitive networks with local graded inhibition; (ii) convergent connections to each neuron from a topologically corresponding region of the preceding layer, leading to an increase in the receptive field size of neurons through the visual processing areas; (iii) synaptic plasticity based on a Hebb-like associative learning rule modified to incorporate a short term memory trace of the preceding activity.

2.3. VisNet architecture

The model consists of a hierarchical series of four layers of competitive networks, corresponding approximately to V2, V4, the posterior inferior temporal cortex (TEO) and the anterior inferior temporal cortex (TE), as shown in Fig. 1. The forward connections to individual cells are derived from a topologically corresponding region of the preceding layer, using a Gaussian distribution of connection probabilities. These distributions are defined by a radius which will contain approximately 67% of the connections from the preceding layer. The values used are given in Fig. 1.

Before stimuli are presented to the network's input layer they are preprocessed by a set of input filters which accord with the general tuning profiles of simple cells in V1. The input filters used are computed by weighting the difference of two Gaussians by a third orthogonal Gaussian according to the following:

$$\Gamma_{xy}(\rho,\theta,f) = \rho\left[e^{-\left(\frac{x\cos\theta+y\sin\theta}{\sqrt{2}/f}\right)^2} - \frac{1}{1.6}e^{-\left(\frac{x\cos\theta+y\sin\theta}{1.6\sqrt{2}/f}\right)^2}\right]e^{-\left(\frac{x\sin\theta-y\cos\theta}{3\sqrt{2}/f}\right)^2}, \quad (1)$$

where f is the filter spatial frequency, θ is the filter orientation, and ρ is the sign of the filter, i.e. ± 1. The shape of these filters has been illustrated previously (Wallis & Rolls, 1997; Rolls & Deco, 2002). Individual filters are tuned to spatial frequency (0.0625–0.5 cycles/pixel); orientation (0° to 135° in steps of 45°); and sign (± 1). The numbers of layer 1 connections to the different spatial frequency filter groups 0.5, 0.25, 0.125, and 0.0625 were 201, 50, 13, and 8, respectively.

The activation h_i of each neuron i in the network is set equal to a linear sum of the inputs x_j from afferent neurons j weighted by the synaptic weights w_{ij}. That is,

$$h_i = \sum_j w_{ij}x_j, \quad (2)$$

where x_j is the firing rate of neuron j, and w_{ij} is the strength of the synapse from neuron j to neuron i.

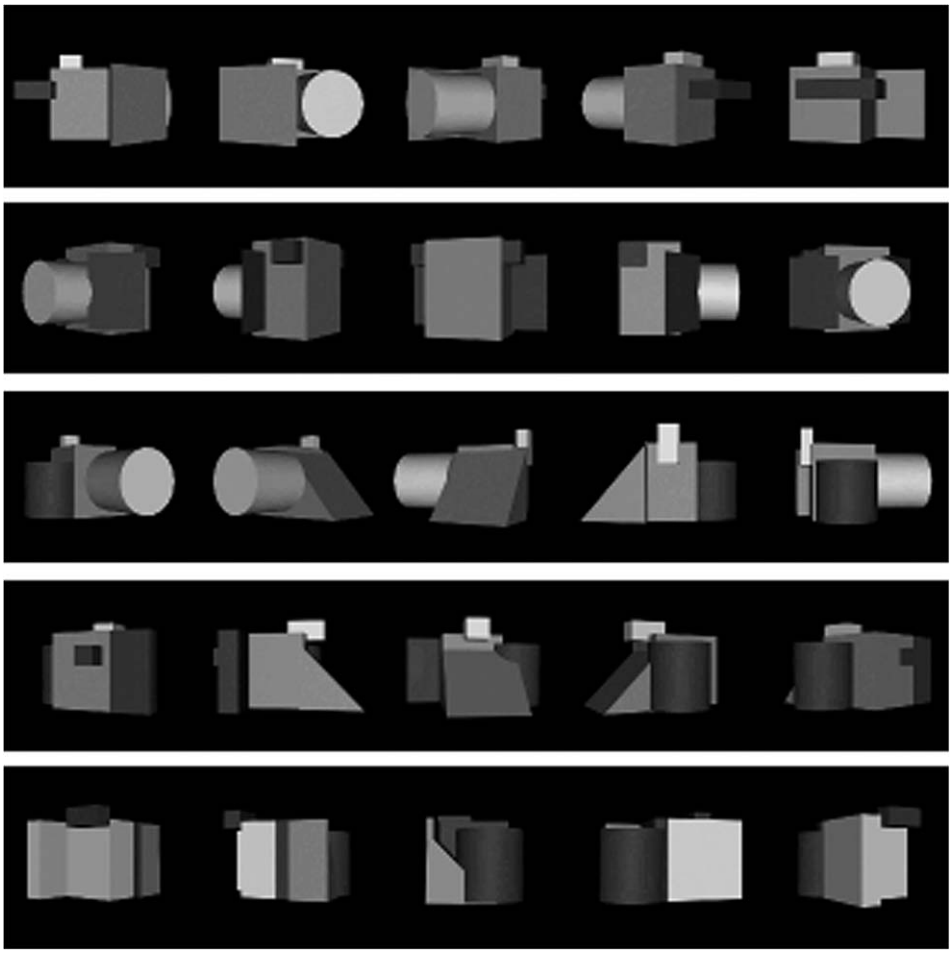


Fig. 3. Views of the objects used in both modeling and human psychophysical experiments. The views shown are from 72° increments around the objects, and hence are the test views used in Experiment 2 (columns show the views that were grouped together during 'interleaved' training in that experiment).

Within each layer competition is graded rather than winner-take-all, and is implemented in two stages. First, to implement lateral inhibition the activations h of neurons within a layer are convolved with a spatial filter, I, where δ controls the contrast and σ controls the width, and a and b index the distance away from the centre of the filter

$$I_{a,b} = \begin{cases} -\delta e^{-\frac{a^2+b^2}{\sigma^2}} & \text{if } a \neq 0 \text{ or } b \neq 0, \\ 1 - \sum_{\substack{a \neq 0 \\ b \neq 0}} I_{a,b} & \text{if } a = 0 \text{ and } b = 0. \end{cases} \quad (3)$$

The lateral inhibition parameter pairs σ and δ for layers 1–4 are, respectively, 1.38 and 1.5; 2.7 and 1.5; 4.0 and 1.6; and 6.0 and 1.4.

Next, contrast enhancement is applied by means of a sigmoid activation function:

$$y = f^{\text{sigmoid}}(r) = \frac{1}{1 + e^{-2\beta(r-\alpha)}}, \quad (4)$$

where r is the activation (or firing rate) after lateral inhibition, y is the firing rate after contrast enhancement, and α and β are the sigmoid threshold and slope, respectively. The parameters α and β are constant within each layer, although α is adjusted to control the sparseness of the firing rates at each timestep. For example, to set the sparseness to, say, 5%, the threshold is set to the value of the 95th percentile point of the activations within the layer. The sigmoid activation function parameter pairs percentile and slope β are for layers 1–4, respectively, 99.2 and 190; 98 and 40; 88 and 75; and 91 and 26.

2.4. Learning rules

One aim of the modeling simulations was to demonstrate the different effects of training produced by both the CT and trace learning approaches, and thus the model was trained separately on these rules, which are specified next.

2.4.1. Continuous transformation (CT) learning

The continuous transformation learning process operates as shown in Fig. 2, and is essentially normal competitive learning (Hertz, Krogh, & Palmer, 1991; Rolls & Treves, 1998; Rolls & Deco, 2002), but operating in a regime with spatial continuity in the input representations. During the presentation of an object at one view that activates particular neurons in the input layer, a small winning set of neurons in the post-synaptic layer will modify (through associative learning) their afferent connections from the input layer to respond well to the object at that view. When the same object appears later at nearby views, so that there is a similarity in spatial form with the trained view, the same neurons in the post-synaptic layer will be activated because some of the active afferents are the same as when the object was trained at the original view, and the synaptic weights from these afferents have been strengthened. The subset of newly active afferents for the new transform will then undergo associative synaptic modification (see Fig. 2). The process can be continued for subsequent shifts in view in order to generate invariance across a large range of views, provided that a sufficient proportion of input cells stay active between individual shifts. A more detailed description of CT learning is provided elsewhere (Stringer et al., 2006).

A variety of associative rules could be used to implement CT learning. In the simulations with CT learning described here we use the following associative (Hebb) learning rule:

$$\delta w_{ij} = \alpha y_i x_j, \tag{5}$$

where δw_{ij} is the increment in the synaptic weight w_{ij}, y_i is the firing rate of the post-synaptic neuron i, x_j is the firing rate of the pre-synaptic neuron j, and α is the learning rate. To bound the growth of each neuron's synaptic weight vector, $\mathbf{w}_i$ for the ith neuron, its length is normalized at the end of each timestep during training as in usual competitive learning (Hertz et al., 1991; Rolls & Deco, 2002).

2.4.2. Trace learning

As outlined in the introduction, trace learning utilizes the temporal continuity of objects in the world in order to help the network develop invariant representations. The trace learning rule (Földiák, 1991; Rolls, 1992; Wallis & Rolls, 1997; Rolls & Milward, 2000; Rolls & Stringer, 2001) encourages neurons to develop invariant responses to input patterns that tend to occur close together in time, because these are likely to be from the same object. The particular rule used (Rolls & Milward, 2000) was

$$\delta w_j = \alpha \bar{y}^{\tau-1} x_j^{\tau}, \tag{6}$$

where the trace $\bar{y}^{\tau}$ is updated according to

$$\bar{y}^{\tau} = (1-\eta) y^{\tau} + \eta \bar{y}^{\tau-1} \tag{7}$$

and we have the following definitions

x_j: jth input to the neuron.
$\bar{y}^{\tau}$: trace value of the output of the neuron at time step τ.
w_j: synaptic weight between jth input and the neuron.
y: output from the neuron.
α: learning rate. Annealed to zero.
η: trace value. The optimal value varies with presentation sequence length.

The parameter η may be set anywhere in the interval [0,1], and for the simulations described here was set to 0.8. Discussions of the good performance of this rule, and its relation to other versions of trace learning rules, are provided elsewhere (Rolls & Milward, 2000; Rolls & Stringer, 2001). Weight normalization is used as described above.

2.5. Training and test procedure

To train the network each stimulus is presented to the network in a sequence of different transforms (e.g. views). At each presentation the activation of individual neurons is calculated, then their firing rates are calculated, and then the synaptic weights are updated. The presentation of all the stimuli across all transforms constitutes 1 epoch of training. In this manner the network is trained one layer at a time starting with layer 1 and finishing with layer 4. In all the investigations described here, the numbers of training epochs for layers 1–4 were 150. The learning rates α in Eqs. (5) and (6) for layers 1–4 were 0.0037, 0.0067, 0.005, and 0.004.

Two measures of performance were used to assess the ability of the output layer of the network to develop neurons that are able to respond with view invariance to individual stimuli or objects (Rolls & Milward, 2000; Stringer et al., 2006).

A single cell information measure was applied to individual cells in layer 4 and measures how much information is available from the response of a single cell about which stimulus was shown independently of view. The measure was the stimulus-specific information or surprise, $I(s,R)$, which is the amount of information the set of responses, R, has about a specific stimulus, s. (The mutual information between the whole set of stimuli S and of responses R is the average across stimuli of this stimulus-specific information.) (Note that r is an individual response from the set of responses R.)

$$I(s,R) = \sum_{r \in R} P(r|s) \log_2 \frac{P(r|s)}{P(r)}. \tag{8}$$

The calculation procedure was identical to that described by Rolls, Treves, Tovee, and Panzeri (1997) with equispaced bins (Rolls & Milward, 2000; Stringer et al., 2006). Because VisNet operates as a form of competitive net to perform categorization of the inputs received, good performance of a neuron will be characterized by large responses to one or a few stimuli regardless of their position on the retina (or other transform such as view), and small responses to the other stimuli. We are thus interested in the maximum amount of information that a neuron provides about any of the stimuli, and this is what is measured by the stimulus-specific information or surprise.

A multiple cell information measure, the average amount of information that is obtained about which stimulus was shown from a single presentation of a stimulus from the responses of many cells, enabled measurement of whether across a population of cells information about every object in the set was provided. Procedures for calculating the multiple cell information measure are given elsewhere (Rolls, Treves, & Tovee, 1997; Rolls & Milward, 2000; Stringer et al., 2006). The multiple cell information measure is the mutual information $I(S,\mathbf{R})$, that is, the average amount of information that is obtained from a single presentation of a stimulus about the set of stimuli S from the responses of all the cells. For multiple cell analysis, the set of responses, $\mathbf{R}$, consists of response vectors comprised by the responses from each cell. Ideally, we would like to calculate

$$I(S,\mathbf{R}) = \sum_{s \in S} P(s) I(s,\mathbf{R}) \tag{9}$$

However, the information cannot be measured directly from the probability table $P(\mathbf{r},s)$ (where the 'stimulus' s refers to an individual object that can occur with different transforms, e.g. translation or size (Wallis & Rolls, 1997), and $\mathbf{r}$ to the response rate vector provided by the firing of the set of neurons to a presentation of that stimulus) because the dimensionality of the response vectors is too large to be adequately sampled by trials. Therefore a decoding procedure is used, in which the stimulus s' that gave rise to the particular firing rate response vector on each trial is estimated. This involves maximum likelihood decoding. For example, given a response vector $\mathbf{r}$ to a single presentation of a stimulus, its similarity to the average response vector of each neuron to each stimulus is used to estimate which stimulus was shown. The probabilities of it being each of the stimuli can be estimated in this way (Rolls et al., 1997). A probability table is then constructed of the real stimuli s and the decoded stimuli s'. From this probability table, the mutual information is calculated as

$$I(S,S') = \sum_{s,s'} P(s,s') \log_2 \frac{P(s,s')}{P(s)P(s')} \tag{10}$$

The multiple cell information was calculated using the five cells for each stimulus with high single cell information values for that stimulus. Thus 25 cells were used in the multiple cell information analysis. The maximum information that can be provided about the stimulus set is

$$\text{Maximum information} = \log_2(\text{Number of stimuli}), \tag{11}$$

where in this case the number of stimuli is 5. This gives a maximum information value of 2.32 bits for both the multiple and single cell information.

2.6. Psychophysics

To test the effects of the spatial and temporal correlations on invariance learning in humans each participant took part in a test session in which they completed a same/different (delayed match to sample) task designed to test their ability to recognise the objects invariantly with respect to view, and discriminate between views of different objects. This initial session established a pre-training baseline of performance for each subject. They then took part in one of the four types of training session outlined above (either interleaved, sequential or permuted training with the experimental objects or interleaved training with the control objects). Finally, they then repeated the test session in order to determine what effects, if any, the training session had on performance.

2.7. Test sessions

Participants were required in the delayed match to sample paradigm to judge, by pressing one of two keys, whether the two test images were of the same or different objects. Each trial was as follows. A fixation cross was displayed centrally for 2000 ms, then the first test image (the 'sample') was presented for 750 ms followed by a 500 ms mask. The screen then remained blank for 250 ms before the second test image (the 'match/non-match') was presented for 750 ms followed by another 500 ms mask. The screen then remained blank until a response button was pressed. The purpose of the mask was to prevent direct image comparison between the sample and match stimuli. Masks were produced by copying randomly chosen 64×64 pixel regions from the test images and pasting them together to create a 256×256 pixel 'patchwork'. Four different masks were created and selected randomly at each image presentation. The relatively short presentation times of the test images in conjunction with a backward masking image were used in order to minimise the opportunity for subjects to use 'cognitive' strategies, such as mental rotation or memorising a verbal description of the object. Test (and training) sessions where implemented and run using Neurobehavioral Systems' (Albany, CA) Presentation software package (version 0.76).

In order to make the experiments as consistent as possible with the VisNet simulations used to generate the experimental predictions, the same set of test images was used. This consisted of each of the five objects seen from five views, each separated by 21.6° in Experiment 1 and by 72° (as illustrated in Fig. 3) in Experiment 2. On match trials, the sample stimulus was one view of an object, and the match stimulus was any other view of the same object. Given that there were five objects, and five views of each object, this produced fifty match trials per test session. (One was the pre-training, and the other was the post-training, test session.)

For non-match trials, the sample stimulus was one view of an object, and the match stimulus was the corresponding view of any other object. (The corresponding views of different objects in Experiment 2 are in the columns of Fig. 3.) This meant that views of two objects that had been close together in time during the interleaved training were used on the same trial in the test session, thus allowing any learned association between (corresponding) views of different objects to be revealed in the test session by false matches on these non-match trials. This produced 50 non-match trials.

The order of trials was randomized within each session. Prior to the first session participants were allowed six practice trials (with visual feedback after each trial) using images from the control set (see below) in order to familiarize themselves with the experimental set up.

2.8. Training session

In psychophysical Experiments 1 and 2, the training conditions were sequential, interleaved, and control. In Experiment 2, the permuted training condition was added to reveal more detail about order effects in the efficacy of learning when temporal correlations within an object are present. A separate group of 10 subjects was used for each condition in each experiment. The subjects in the control condition saw the same objects as the experimental subjects during testing, but the 'training' objects were a different set of objects not used in the testing sessions. During training each image appeared for 750 ms, with no gaps between images. The images used were 25 views, each separated by 3.6° in Experiment 1 and (as explained in Section 3) by 14.4° in Experiment 2. (In the sequential and permuted conditions, the sequence of views for an individual object was shown twice before the next object appeared so that views at the end of each sequence would have the chance to be temporally associated with views at the start of the sequence. In the interleaved and control conditions the entire training sequence repeated itself, in order to make sure that in all four training conditions every view was seen the same number of times (twice) during a session.)

2.9. Stimulus presentation and participants

During both test and training sessions all images were displayed on a black background using a CRT monitor with a resolution of 1280×1024 and a refresh rate of 75 Hz. Participants were seated level with the screen and approximately 40 cm away. Stimuli varied in size between 2° and 4° horizontally, depending on the particular viewing angle of the object, and around 2° vertically (with minor variations of $<0.5°$ depending on the specific view). The instructions given to the subjects were to press one key if the second stimulus on a trial was the same object as the first stimulus, and to press a different key if the two images on a single trial were of two different objects. The instructions for the training period were to watch the images on the screen carefully as this might help them to perform better in the second testing session.

Thirty participants in Experiment 1 and 40 participants in Experiment 2 (ten per training group) were tested in total. All were first year psychology undergraduates who participated in exchange for course credit. All had normal or corrected to normal vision.

3. Results

3.1. Modeling

The simulations described in this section investigate whether continuous transformation learning which uses a purely associative learning rule can learn view invariant representations of multiple (five) objects, each of which is a complex object that consists of 5–6 shape primitives (see Fig. 3). (The initial demonstration of CT learning was with only two objects each composed of a single shape primitive (Stringer et al., 2006)). The simulations then analyse the properties of the CT learning with training conditions that include interleaved vs sequential training, and different spacing between the different views, and compare it to temporal trace learning under the same conditions. In addition to extending our understanding of continuous transformation learning in this way, the simulations in this section also address how spatial and temporal continuity are important in networks that learn invariant representations with exactly the same set of stimuli used in the psychophysical experiments. This provides a basis for interpreting the mechanisms that may underlie the psychophysical findings on humans' learning of invariant representations of complex but well-defined objects. Key predictions of the two theories are tested, including the predictions that CT but not temporal trace learning can self-organize invariant representations when the views of different objects are interleaved, and that CT learning but not necessarily temporal trace learning will perform poorly if the spacing between the closest views become larger, thus breaking the spatial continuity in the images seen by the network (see Section 2).

3.2. Properties of CT and trace learning with several complex objects

The results of training the network with an associative (Hebb) rule in the CT paradigm, with a temporal trace rule (see Eq. (6) in Section 2), and, for comparison,

without training are shown in Fig. 4. This figure shows the average performance over five simulation runs of the network when trained with different numbers of views of the stimuli (each separated by 3.6°) to assess the performance as the network is loaded with more views, together subtending larger viewing angles, when trained with CT learning. Performance was measured from the single cell information analysis by the number of layer 4 cells that have the maximum information of 2.32 bits, and in the multiple cell analysis by the maximum value of the information reached. The results show that the CT paradigm performs well, and can form view invariant representations of multiple complex objects with only a purely associative learning rule.

The results in Fig. 4 also show that the trace rule performs somewhat better (and without training there are almost no single cells that perform perfectly). The finding that the learning in the Hebb (CT) condition is quite good indicates that continuous transformation learning may contribute to the good performance of learning with a trace rule. This may occur because, during learning with successively presented transforms of the same object (as used in this and the following simulation), the CT effect tends to keep neurons in higher layers active across successive transforms of the object, and this in turn leads to high trace values $\overline{y}^{\tau}$ for the active neurons in the higher layers by virtue of equation (7).

To analyse how the spacing between views influences the performance with the two rules (and we predict that at some point Hebb learning will start to break down as the CT effect can no longer operate because the adjacent views are too different), we show in Fig. 5 the operation when the total number of views is fixed (at 10), but the spacing between each view is varied. It is shown that CT learning does perform less well than trace learning as the spacing between views is increased.

3.3. Simulations corresponding to the psychophysical experiments

We next used both Hebb (CT) and trace learning to train the network with the two presentation methods used in both psychophysical experiments, in order to compare the performance of the model using the two different learning rules with the human experimental data. In the first, 'sequential', presentation method, all training views of the first object were presented in sequential order of clockwise rotation, followed by all views of the second object, etc. In the second 'interleaved', method, the training sequence consisted in the views of the different objects interleaved with one another as they rotated (see Section 2). It was predicted that CT learning should continue to produce successful learning with this interleaved paradigm, whereas the performance of trace learning should be impaired.

As in the human psychophysics, the network's performance was based on a same/different comparison for whether two views were of the same object (from different viewing angles), or from different objects, based on the stimulus indicated by the multiple cell analysis when each test view was shown to the network. (The multiple cell information analysis described in Section 2 makes a prediction from the responses of multiple cells in layer 4 about which object was shown on each trial.) The 100 trials (50 match, 50 non-match) corresponded to those used in the psychophysics.

Fig. 6a shows the results corresponding to the conditions of psychophysical Experiment 1, with the post-training results shown as session 2. With sequential training, networks trained with either the trace or the Hebb learning rule perform equally well. With interleaved training, the Hebb-trained (CT) network continues to perform well. However, as predicted, with interleaved training, the trace rule trained network performs poorly (and indeed worse

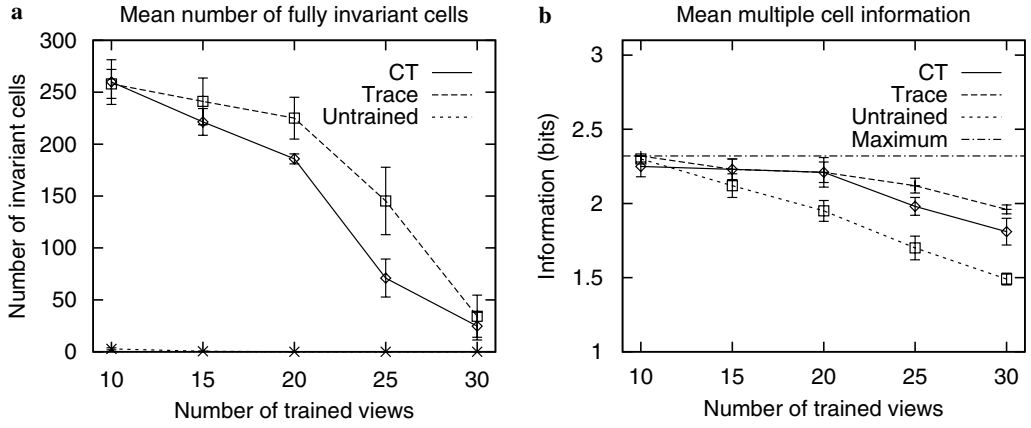


Fig. 4. (a) Mean number of cells with the maximum single cell information (2.32 bits) in the output layers of five networks trained with varying numbers of views and using different methods of learning (CT, trace or no training). Each network was trained using a different set of views from the same five objects. Successive views in each sequence were separated by 3.6°. (b) The mean multiple cell information measured from the same set of networks. The horizontal line at 2.32 bits labelled maximum indicates the amount of information required to perfectly discriminate the set of 5 objects, and therefore the maximum that could be made available by the cell population.

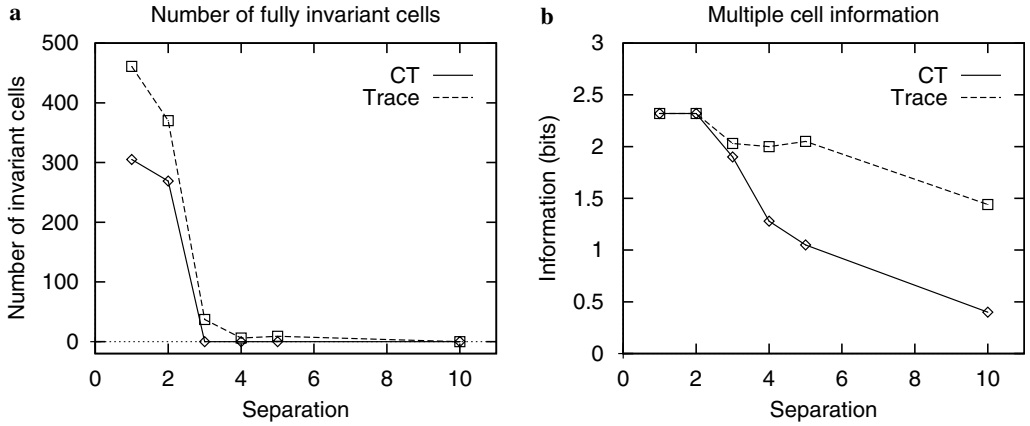


Fig. 5. (a) The number of cells with the maximum single cell information (2.32 bits) in the output layer when the network was trained with 10 views. In different simulations, the angle separating each view was varied. A view separation of 1 corresponds to 3.6° between views, and a view separation of 10 corresponds to 36° between views. The network was trained with the Hebb or the trace rule. The network was trained with five objects. (b) The corresponding multiple cell information measures.

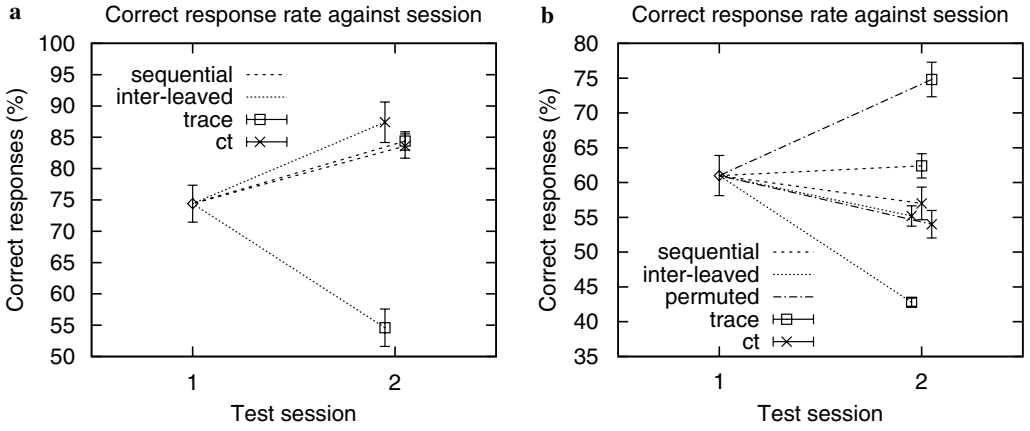


Fig. 6. (a) Simulations corresponding to psychophysical Experiment 1 showing the percentage correct (mean ± SE across five different simulation runs with different random seeds) for test session 1 (before training) and test session 2 (after training) for four different training conditions: sequential presentation with the Hebb (CT) rule; sequential presentation with the trace rule; interleaved presentation with the Hebb (CT) rule; and interleaved presentation with the trace rule. Five objects were used, with 25 training views spaced 3.6° apart, and 5 test views spaced 18° apart. (b) Simulations corresponding to psychophysical Experiment 2 showing the percentage correct (mean ± SE) for test session 1 (before training) and test session 2 (after training) for six different training conditions: sequential presentation with the Hebb (CT) rule; sequential presentation with the trace rule; interleaved presentation with the Hebb (CT) rule; interleaved presentation with the trace rule; permuted presentation with the Hebb (CT) rule; and permuted presentation with the trace rule. Five objects were used, with 25 training views spaced 14.4° apart, and 5 test views spaced 72° apart.

than the untrained network the performance of which is shown as session 1), because views of different objects become associated together, impairing invariant object recognition. An ANOVA showed that (as predicted) the performance of both the CT conditions, and the sequential condition with trace, performed better than without training ($p < .05$ one-tailed in all cases); and that the interleaved training with the trace rule performed as predicted less well than the untrained network ($F(1,4) = 14.39$, $p < .01$ one-tailed).

We also simulated training conditions that correspond to those used in psychophysical Experiment 2, which used a different set of test and training views (see psychophysics Experiment 2) and included a new permuted training condition in which the views of a given object were presented in permuted rather than sequential order. Comparison of sequential with permuted training provides evidence on whether regularity in the way in which an object transforms is important for learning invariant representations. Fig. 6b shows that the best performance was with the permuted views and trace rule training. This protocol enabled different pairs of views of a given object to be associated together because different pairs of views could be close together in time and thus temporally associated. The trace rule performed less well with sequential training, consistent with the fact that very different views of an object (which spanned 360° in this simulation) occurred distantly in time, so that the temporal trace rule could not enable them to be associated. As in the previous simulation, and as predicted, in the interleaved condition the trace rule produced worse

performance than without training, because views of different objects were being associated together. All the CT conditions performed relatively poorly in these simulations, probably because the views were widely separated (by 14.4°), so that the CT effect could not operate, at least in a network with no previous training. An ANOVA showed that the permuted trace rule condition performed better as predicted than without training (e.g. $F(1,4) = 8.16$, $p < .025$); and that the interleaved training with the trace rule performed less well than the untrained network ($F(1,4) = 47.05$, $p < .001$ one-tailed).

3.4. Psychophysics

In order to test the effects of the spatial and temporal correlations on invariance learning in humans, each participant took part in a test session in which they completed a same/different (delayed match to sample) task designed to test their ability to recognise the objects invariantly with respect to view, and discriminate between views of different objects. This initial session established a pre-training baseline of performance for each subject. They then took part in one of the different types of training session described in Section 2 (either interleaved, sequential or permuted training with the experimental objects, or for control subjects interleaved training with control objects that were not from the test set, though were constructed similarly). After training, participants then repeated the test session in order to determine what effects, if any, the training session had on performance.

3.5. Psychophysics Experiment 1

The images used were 25 views, each separated by 3.6°, and the training conditions were sequential, interleaved or control. Fig. 7 shows the mean percentage of correct responses averaged across subjects within a group for each test session. The ANOVA (with test session as a within subjects factor and training condition as a between subjects factor) revealed a significant interaction effect showing differences between the different training conditions ($F(2,23) = 4.27$, $p = .027$). Post hoc (simple main effects) tests showed in the sequential training group that the improvement in performance was highly significant ($F(1,23) = 17.61$, $p = .001$), and in the interleaved group was also significant ($F(1,23) = 7.37$, $p = .012$). There was no significant effect in the control group which had been 'trained' with a different set of objects to those used in testing ($F(1,23) = .054$, $p = .82$). (The data shown were from match trials, that is trials on which the participants judged whether the sample and match were different views of the same object, as these trials directly test whether subjects can recognise that two views are of the same object, that is perform invariant object recognition. The same pattern of results was obtained if data from match and non-match trials were included in the analysis, though with a slightly reduced level of significance, and as the training effects in this experiment were more clearly evident in the match trials, these are the data used to illustrate the findings. Four subjects were not included in this analysis as their performance in the pretraining period was greater than 90% correct, leaving little room for improvement as a result of training to be demonstrated.)

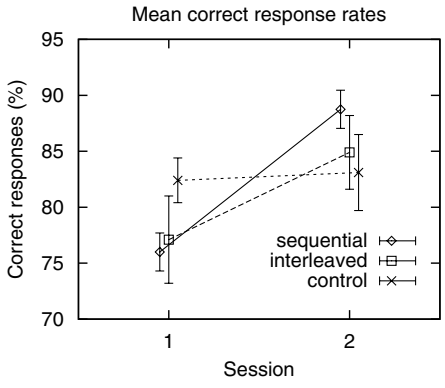


Fig. 7. Psychophysical Experiment 1. Mean percentage correct responses, averaged across each subject within a training group, ±SEM. Session 1 is pretraining, and session 2 is post-training.

The improvement in performance in the sequential condition is consistent with the hypothesis that trace and/or continuous transformation learning are involved. The improvement in performance in the interleaved condition is consistent with the hypothesis that continuous transformation learning is involved, and indeed provides strong support for a role of spatial continuity that is independent of temporal continuity in human view invariance learning.

However, subjects were, on average, correct in 78.6% of trials in the first test session. As the performance before training was well above chance, the possibility for strong further improvement may have limited the sensitivity of Experiment 1. We therefore performed Experiment 2 in which the task was made more difficult, thus potentially allowing further learning effects to be revealed, and also allowing a new training condition with permuted sequences to be investigated.

3.6. Psychophysics Experiment 2

In Experiment 1, the angle between the most distant test (and training) views of an object was 90°. This meant that in match trials the angle between test views was never greater than this value and was often smaller—as little as 21.6°. This may in part explain why subjects made relatively few errors in Experiment 1. In Experiment 2 the task was made more difficult by using test views that were separated by larger angles of separation, 72°, between views, which covered the full 360° extent. Fig. 3 shows the set of test views. The training views were separated by 14.4°. A new group of subjects was used, with 10 subjects per group, and a permuted group was added.

Fig. 8 shows the mean percentage of correct responses averaged across subjects within a group for each test

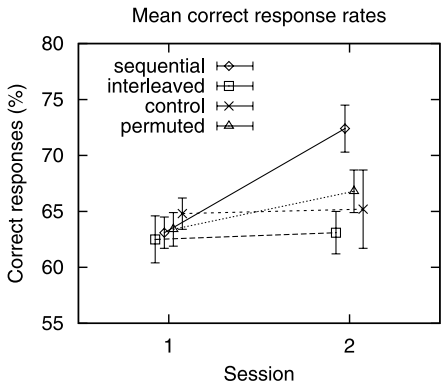


Fig. 8. Psychophysical Experiment 2. Mean percentage correct responses, averaged across each subject within a training group, ±SEM. Session 1 is pretraining, and session 2 is post-training.

session. The use of the new views had the expected effect of decreasing initial performance, with subjects making correct responses on 63.5% of trials on average in the first session (with 50% being the chance level of performance).

An ANOVA (with test session as a within subjects factor and training condition as a between subjects factor) revealed a very significant interaction effect showing differences between the different training conditions ($F(3,36) = 4.92$, $p = .006$). Fig. 8 shows that improvements were not consistent across all four conditions. Post hoc (simple main effects) tests showed in the sequential training group that the improvement in performance was highly significant ($F(1,36) = 24.73$, $p < .0001$). In the interleaved group the improvement was not significant ($F(1,36) = 0.10$, $p = .75$). In the permuted group the improvement was not significant ($F(1,36) = 3.31$, $p = .08$). There was no significant effect in the control group which had been 'trained' with a different set of objects ($F(1,36) = 0.05$, $p = .83$).

A preplanned comparison showed that there was a significant interaction between the performance of the sequential and the interleaved group ($F(1,18) = 10.60$, $p = .004$), with the sequential training condition producing better results. This is consistent with the hypothesis that trace learning is being used, and more generally that temporal contiguity of the different transforms of objects is useful in invariance learning. The poorer performance in the interleaved condition is an indication that trace learning is more efficacious than continuous transform learning, at least in this experiment, although continuous transformation learning could be at a disadvantage here because the difference between the different training views is relatively large (14.4°).

Another preplanned comparison showed that there was a significant interaction between the performance of the sequential and the permuted group ($F(1,18) = 6.63$, $p = .019$), with the sequential training condition producing better results. This is consistent with the hypothesis that temporal contiguity of the different transforms of objects that are close is useful in invariance learning. Effectively, if subjects saw an object rotating in the correct order of views rather than jumping randomly between the different training views, this enabled better learning. The humans, and the learning algorithm in the brain, thus perform better in a condition when it may be possible to link together different nearby views, rather than making random associations between views.

We note that just seeing one object for a long period before another object is shown does not appear to be the critical training variable, for performance was better in the sequential than the permuted condition, even though each object was seen for the same continuously long period in each condition. Instead, the smooth transition between views of an object, which did distinguish the sequential from the permuted training condition, was a factor that improved learning (see Section 4).

We note that one hypothesis that learning in the sequential condition might be better than in the interleaved condition just because a single object is available for a long period before another one is shown is not supported by the finding in this experiment that performance with the sequential condition is much better than in the permuted condition.

4. Discussion

The modeling results provide evidence that a new principle of invariance learning, continuous transformation learning, is an important learning principle in neuronal networks that can help them to learn invariant representations of complex objects. The principle is important in the sense that networks with no temporal trace learning but only purely associative (Hebbian) learning can use continuous transformation learning to learn invariant representations. The psychophysical results show that some invariant object recognition learning can occur in a condition with spatial but not temporal continuity, and this is consistent with the hypothesis that a continuous transformation learning or a similar process can contribute to human view invariance learning. At the same time, the results show that temporal associations are also a very important factor in view invariant object learning.

The modeling results are new, in that they show that CT learning (which uses a purely associative synaptic modification rule) is capable of good invariance learning when five objects are trained simultaneously, and when each object is spatially quite complex, consisting of a combination of 5–6 3D shape primitives. These results are shown in Fig. 4.

The results in Fig. 4 also show that the trace learning rule performs a little better than the Hebb (CT) learning condition when the different views of an object are presented without interleaving. Moreover the results in Fig. 5 show that as the spacing between the views is increased the Hebb (CT) learning starts to break down as the CT effect can no longer operate because the adjacent views are too different. This prevents the Hebb rule from developing invariant neurons, and degrades the performance

with the trace rule which must now rely on temporal continuity alone, with no contribution of spatial continuity.

The results of the interleaved training shown in Fig. 6a indicate that the trace rule trained network performs poorly (and indeed worse than the untrained network), because views of different objects become associated together, impairing invariant object recognition. CT learning is not impaired by the interleaved object presentations, because it uses spatial similarity to drive the learning.

The psychophysics showed in Experiment 1 that there was some view invariant learning when the views of the five objects were interleaved (see Fig. 7). This is evidence that a learning mechanism for view invariance operates that is of the type predicted by continuous transformation learning, as learning occurs in the absence of close temporal correlations between different views of the same object. Further evidence that some CT-like learning applies in humans is that when the spacing between the views was increased from 3.6° in Experiment 1 to 14.4° in Experiment 2, then view invariant learning was no longer evident in the interleaved condition. This is predicted by CT learning, as when the spacing between views becomes too large, there is insufficient spatial continuity to ensure that some of the same neurons are activated by adjacent stimuli. Consistent with this, in the simulations shown in Fig. 5, CT learning occurred when the separation between views of the same objects was 3.6° or 7.2°, but was not present when the separation was 18° or more. Thus, the results described in this paper show that in humans some invariance learning can occur when temporal correlations do not provide a learning cue, and we propose that a mechanism that underlies such view invariance learning is continuous transformation learning. This hypothesis has not been tested, to our knowledge, in previous psychophysical investigations.

The psychophysics also showed that when the spatial continuity between successive views is decreased in Experiment 2 in the interleaved condition by using a wider (14.4°) separation between views, nevertheless there was very clear view invariance learning in the sequential condition (see Fig. 8). This indicates that temporal contiguity of views of the same object is important in human view invariance learning. Correspondingly, in the simulations shown in Fig. 5 (right), when the view separation was large (18° or more), trace learning performed better than CT learning. (We note that the failure to improve performance in the interleaved training condition shows that just increasing familiarity with the views was not sufficient to improve performance.)

Interestingly, sequential training in psychophysics Experiment 2 produced better performance than training with permuted views within an object. The implication for invariance learning is that temporal associativity can be more effective when the transition between views is consistent with the way in which the object would transform in the real world. (For example, objects normally rotate continuously, and do not jump suddenly through large viewing angles, or with great regularity backwards and forwards between two different objects.) This is consistent with the general principle of trace learning being an important mechanism of view invariance learning, but suggests that spatial continuity between views can be a useful additional constraint. One possibility is that if there is an obvious spatial discontinuity in the set of images, the temporal trace learning mechanism may be reset, by for example inhibiting neuronal firing that might provide a representation of previous neuronal activity. The results also establish that view invariance learning using this temporal continuity can be very fast, in that in the psychophysical experiments, the training consisted effectively of two trials. (That is, during training, each view of an object was shown twice.)

Trace learning predicts that, in the absence of other constraints, associations may be made between different objects if they are interleaved during training, and this effect was observed in our simulations in that in the trace learning condition learning was impaired with interleaved stimuli (see Fig. 6). In our psychophysical experiments, there was no decrease in performance in the interleaved condition, and this implies that any possible disadvantage of temporal continuity in producing spurious associations between interleaved objects was small. Some evidence for spurious associations when views of different faces are interleaved has been found (Wallis & Bülthoff, 2001; Wallis, 2002) if the interleaved views were presented in such a way that it appeared that a single face was rotating, with the views in sequential order. That evidence, obtained with interleaved faces and spurious associations, is thus consistent with the finding described here that when view invariant representations of objects are learned, temporal associations are important, and operate best when an object transforms in a manner consistent with the sequence of views that would be generated when the object transforms in the real world, for example, the sequential views produced when it rotates (see our psychophysical Experiment 2).

An important implication of the results shown here is that networks can learn invariant representations of objects without using temporal contiguity. An important learning principle that underlies such learning uncovered in this and a related paper (Stringer et al., 2006) is close spatial similarity between the training images. Given that at the behavioral level some view invariant learning without temporal contiguity has been found in monkeys (Wang, Obama, Yamashita, Sugihara, & Tanaka, 2005), an effect we describe for humans in Fig. 7, we note and indeed propose that a possible underlying mechanism for both sets of results is the continuous transformation learning that we describe. We further propose that continuous transformation learning could contribute to learning in any topologically mapped sensory system in the brain, including the somatosensory system. Indeed, CT learning is very likely to operate in such systems under certain parameters, including a synaptic learning rate that is sufficiently high, and continuous variation of the sensory input.

Although we did not observe CT-like learning in humans with the larger separations between views used in psychophysical Experiment 2, we believe that it would be interesting to explore this further. CT-like learning might operate even if during training there is some considerable spacing between views, as a result of prior experience in which humans have learned the low-level spatial statistics that are common to different objects. In these circumstances, only the higher levels of the system need to learn the high-level spatial similarity that defines a new object. By higher levels we mean those areas that are equivalent to high levels of a feature hierarchy network with convergent feedforward connections, as is modeled by VisNet, e.g. areas TE and TEO. Indeed, just this effect has been demonstrated in a previous investigation, in which early layers of the network were trained on features, and then only training in higher levels of the network was necessary to learn a view invariant representation of a new object when the different views were widely spaced (Stringer et al., 2006). Generalization to new untrained views of objects has also been demonstrated in VisNet (Stringer & Rolls, 2002). The proposal we thus make is that generalization to new transforms after training with only one of a few transforms of a new object is likely to improve the more the system (whether animal or network) is trained on increasing numbers of different objects, for then the invariant representations already learned for parts of other objects can help the new object to be recognized from previously unseen transforms. This capacity is well developed in humans, and non-human primates (Wang et al., 2005), and though also a property of VisNet (Stringer & Rolls, 2002; Stringer et al., 2006), it is predicted that this capability will become more prominent in networks such as VisNet as the network is scaled up and trained on larger numbers of objects. At present VisNet is a relatively small model with 1024 neurons per layer, and is designed to investigate the principles of invariant object recognition. Because it is relatively small, we cannot represent in it all the features and sub-parts of objects that humans and animals can represent, and on which extensive training is received with large numbers of objects over many years. It will therefore be of interest in future to test how well it does scale up. We also note that a human strategy might be to pay attention to one feature or part of an object to help with invariant object recognition, and this strategy is not open to this version of VisNet, which is a feedforward network.

Nevertheless, an important conclusion of the present results is that temporal contiguity helps the performance to be better than with purely associative Hebbian learning that applies in CT learning. One reason for this is that temporal contiguity, a property of the transforms of real objects in the world, can help to break the associativity implied by CT learning when this is not appropriate for defining an object. Indeed, a danger of CT learning is that some images of different objects might be sufficiently similar that two different objects become associated. The lack of temporal contiguity between the different objects can, in this case, help to break apart the representations of those objects. In addition, if an object shows a catastrophic view change, as when the cusp of a cup allows the inside of the cup to come into view, there might be insufficient spatial continuity to drive the CT learning. In this case, temporal contiguity, and its instantiation by for example a trace learning rule, can help to associate together spatially dissimilar views of the same object.

In conclusion, the results show that continuous transformation learning is an important principle of training networks to develop invariant representations of even complex objects, and that it may contribute to human invariant learning which can occur with interleaved views of different objects, which breaks the temporal continuity within an object (psychophysics Experiment 1). However, human view invariant object recognition can be better when temporal continuity is present as in the sequential condition in psychophysics Experiment 2, and this is consistent with the hypothesis that temporal trace learning can contribute to invariant object recognition.

Acknowledgment

This research was supported by the Wellcome Trust, and by the Medical Research Council.

References

Bartlett, M. S., & Sejnowski, T. J. (1998). Learning viewpoint-invariant face representations from visual experience in an attractor network. *Network: Computation in Neural Systems, 9*, 399–417.

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

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review, 94*(2), 115–147.

Biederman, I., & Bar, M. (1999). One-shot viewpoint invariance in matching novel objects. *Vision Research, 39*, 2885–2899.

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 1162–1182.

Booth, M. C. A., & Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex, 8*, 510–523.

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience, 3*, 1–8.

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

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

Hasselmo, M. E., Rolls, E. T., Baylis, G. C., & Nalwa, V. (1989). Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research, 75*, 417–429.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. Wokingham, UK: Addison Wesley.

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

Kobotake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology, 71*, 856–867.

Op de Beeck, H., & Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. *Journal of Comparative Neurology, 426*, 505–518.

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

Rolls, E. T. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philosophical Transactions of the Royal Society, 335*, 11–21.

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

Rolls, E. T. (2006). The representation of information about faces in the temporal and frontal lobes of primates including humans. *Neuropsychologia* (in press), doi:10.1016/j.neuropsychologia.2006.04.019.

Rolls, E. T., Aggelopoulos, N. C., & Zheng, F. (2003). The receptive fields of inferior temporal cortex neurons in natural scenes. *Journal of Neuroscience, 23*, 339–348.

Rolls, E. T., & Baylis, G. C. (1986). Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Experimental Brain Research, 65*, 38–48.

Rolls, E. T., Baylis, G. C., & Hasselmo, M. E. (1987). The responses of neurons in the cortex in the superior temporal sulcus of the monkey to band-pass spatial frequency filtered faces. *Vision Research, 27*, 311–326.

Rolls, E. T., Baylis, G. C., & Leonard, C. M. (1985). Role of low and high spatial frequencies in the face-selective responses of neurons in the cortex in the superior temporal sulcus. *Vision Research, 25*, 1021–1035.

Rolls, E. T., & Deco, G. (2002). *Computational neuroscience of vision*. Oxford: Oxford University Press.

Rolls, E. T., & 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*, 2547–2572.

Rolls, E. T., & Stringer, S. M. (2001). Invariant object recognition in the visual system with error correction and temporal difference learning. *Network: Computation in Neural Systems, 12*, 111–129.

Rolls, E. T., & Treves, A. (1998). *Neural networks and brain function*. Oxford: Oxford University Press.

Rolls, E. T., Treves, A., Tovee, M., & Panzeri, S. (1997). Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. *Journal of Computational Neuroscience, 4*, 309–333.

Rolls, E. T., Treves, A., & Tovee, M. J. (1997). The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *Experimental Brain Research, 114*, 149–162.

Stone, J. V. (1996). Learning perceptually salient visual parameters using spatiotemporal smoothness constraints. *Neural Computation, 8*, 1463–1492.

Stringer, S. M., Perry, G., Rolls, E. T., & Proske, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biological Cybernetics, 94*, 128–142.

Stringer, S. M., & Rolls, E. T. (2002). Invariant object recognition in the visual system with novel views of 3D objects. *Neural Computation, 14*, 2585–2596.

Tanaka, K., Saito, H., Fukada, Y., & Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology, 66*, 170–189.

Tovee, M. J., Rolls, E. T., & Azzopardi, P. (1994). Translation invariance and the responses of neurons in the temporal visual cortical areas of primates. *Journal of Neurophysiology, 72*, 1049–1060.

Ullman, S. (1996). *High-level vision*. Cambridge, MA: MIT Press.

Vogels, R., & Biederman, I. (2002). Effects of illumination intensity and direction on object coding in macaque inferior temporal cortex. *Cerebral Cortex, 12*, 756–766.

Wallis, G. (2002). The role of object motion in forging long-term representations of objects. *Visual Cognition, 9*, 233–247.

Wallis, G., & Bülthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences, 98*, 4800–4804.

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

Wang, G., Obama, S., Yamashita, W., Sugihara, T., & Tanaka, K. (2005). Prior experience of rotation is not required for recognizing objects seen from different angles. *Nature Neuroscience, 8*, 1768–1775.

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Computation, 14*, 715–770.