Biol Cybern (2006) 94: 128–142
DOI 10.1007/s00422-005-0030-z

ORIGINAL PAPER

Biological Cybernetics

S. M. Stringer · G. Perry · E. T. Rolls · J. H. Proske

Learning invariant object recognition in the visual system with continuous transformations

Received: 26 July 2005 / Accepted: 5 October 2005 / Published online: 21 December 2005

Abstract The cerebral cortex utilizes spatiotemporal continuity in the world to help build invariant representations. In vision, these might be representations of objects. The temporal continuity typical of objects has been used in an associative learning rule with a short-term memory trace to help build invariant object representations. In this paper, we show that spatial continuity can also provide a basis for helping a system to self-organize invariant representations. We introduce a new learning paradigm "continuous transformation learning" which operates by mapping spatially similar input patterns to the same postsynaptic neurons in a competitive learning system. As the inputs move through the space of possible continuous transforms (e.g. translation, rotation, etc.), the active synapses are modified onto the set of postsynaptic neurons. Because other transforms of the same stimulus overlap with previously learned exemplars, a common set of postsynaptic neurons is activated by the new transforms, and learning of the new active inputs onto the same postsynaptic neurons is facilitated. We demonstrate that a hierarchical model of cortical processing in the ventral visual system can be trained with continuous transform learning, and highlight differences in the learning of invariant representations to those achieved by trace learning.

1 Introduction

There is now much evidence demonstrating that over successive stages the visual system develops neurons that respond with view, size and position (translation) invariance to objects or faces (Rolls 1992, 2000; Rolls and Deco 2002; Desimone 1991; Tanaka et al. 1991). For example, it has been shown that the inferior temporal visual cortex has neurons that respond to faces and objects with translation (Op de Beeck and Vogels 2000; Kobotake and Tanaka 1994; Ito et al. 1995; Tovee et al. 1994), size (Rolls and Baylis 1986; Ito et al. 1995), contrast (Rolls and Baylis 1986), lighting (Vogels and Biederman 2002), spatial frequency (Rolls et al. 1985, 1987), and view (Hasselmo et al. 1989; Booth and Rolls 1998) invariance. It is crucially important that the visual system builds invariant representations, for only then one-trial learning about an object can generalize usefully to other transforms of the same object (Rolls 2005; Rolls and 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 and Poggio 1999; Biederman 1987; Ullman 1996; Rolls and Deco 2002; Elliffe et al. 2002).

In this paper we describe a quite general learning principle, *continuous transformation* (CT) *learning,* that could be used in several sensory systems to build invariant representations. CT learning utilizes spatial continuity of objects in the world. The system we describe is quite powerful, for it relies on spatial overlap between stimuli in small regions of the space, but enables transforms in quite distant parts of the continuous space to be associated together onto the same population of postsynaptic neurons. We show how CT learning could be used in the hierarchical processing that is a property of cortical architecture, in which key principles agreed by many investigators (Riesenhuber and Poggio 1999; Fukushima 1980; Wallis and Rolls 1997) include feedforward connectivity, local lateral inhibition within a layer to implement competition, and then some form of associative learning. Then we show by simulation how it can be used to build invariant representations in a hierarchical network model (VisNet) of cortical processing in the ventral visual system, and show how CT learning differs from but could complement a different invariance learning principle, trace learning (Rolls and Milward 2000; Rolls and Stringer 2001; Wallis and Rolls 1997). Other models with hierarchically organized competitive networks designed to study neurally

S. M. Stringer · G. Perry · E. T. Rolls · J. H. Proske
Centre for Computational Neuroscience,
Department of Experimental Psychology,
Oxford University,
South Parks Road, Oxford
OX1 3UD, England
Tel.: +44-1865-271348
Fax: +44-1865-310447,
E-mail: Edmund.Rolls@psy.ox.ac.uk; www.cns.ox.ac.uk

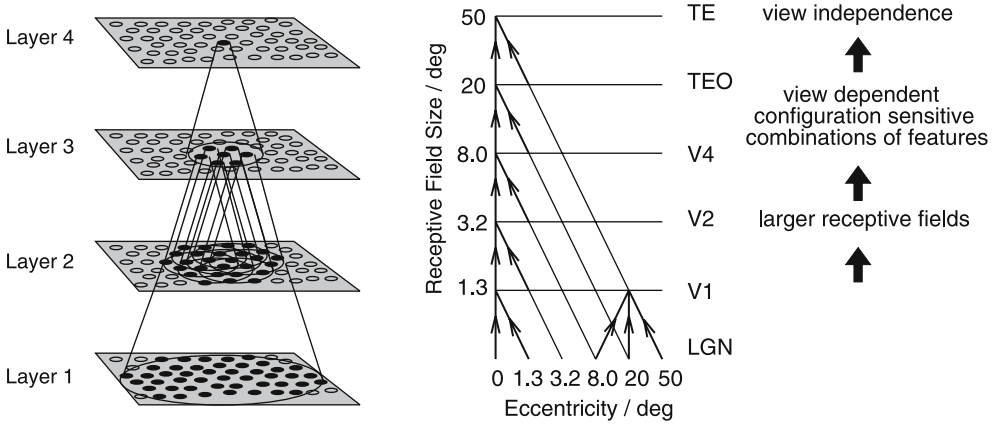

Fig. 1 *Left*: Stylised image of the four layer network. Convergence through the network is designed to provide fourth layer neurons with information from across the entire input retina. *Right*: Convergence in the visual system V1: visual cortex area V1; *TEO* posterior inferior temporal cortex, *TE* inferior temporal cortex (IT)

plausible ways of forming invariant representations of stimuli have been studied by a number of investigators (Riesenhuber and Poggio 1999; Fukushima 1980).

2 Methods

2.1 The VisNet architecture

The model architecture (VisNet) implemented by Wallis and Rolls (1997) that is used to investigate the properties of CT learning in this paper is based on the following: (a) A series of hierarchical competitive networks with local graded inhibition. (b) 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. (c) Synaptic plasticity based on a Hebb-like learning rule. Model simulations which incorporated these hypotheses with a modified associative learning rule to incorporate a short-term memory trace of previous neuronal activity were shown to be capable of producing stimulus-selective but translation and view invariant representations (Rolls and Milward 2000; Rolls and Stringer 2001; Wallis and Rolls 1997).

In this paper, the new CT learning principle in the model architecture (VisNet) uses only spatial continuity in the input stimuli to drive the Hebbian associative learning with no temporal trace. In principle, the CT learning mechanism we describe could operate in various forms of feedforward neural network, with different forms of associative learning rule or different ways of implementing competition between neurons within each layer.

The model consists of a hierarchical series of four layers of competitive networks, corresponding approximately to V2, V4, the posterior inferior temporal cortex (TEO in Fig. 1), and the anterior inferior temporal cortex (TE in Fig. 1), 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 Table 1.

Before stimuli are presented to the network's input layer they are pre-processed by a set of input filters which accord with the general tuning profiles of simple cells in V1. (These input filters which provide the input to layer 1 of VisNet thus correspond to V1 in the right part of Fig. 1.) The input filters used are computed by weighting the difference of two Gaussians by a third orthogonal Gaussian according to the following:

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

where f is the filter spatial frequency, θ is the filter orientation, and ρ is the sign of the filter, i.e. ± 1. Individual filters are tuned to spatial frequency (0.0625–0.5 cycles/pixel); orientation (0–135° in steps of 45°); and sign (± 1). The number of layer 1 connections to each spatial frequency filter group is given in Table 2.

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

Table 1 Network dimensions showing the number of connections per neuron and the radius in the preceding layer from which 67% are received

	Dimensions	Number of connections	Radius
Layer 4	32 × 32	100	12
Layer 3	32 × 32	100	9
Layer 2	32 × 32	100	6
Layer 1	32 × 32	272	6
Retina	128 × 128 × 32	–	–

Table 2 Layer 1 connectivity

Frequency	0.5	0.25	0.125	0.0625
Number of connections	201	50	13	8

The numbers of connections from each spatial frequency set of filters are shown
The spatial frequency is in cycles per pixel

$$h_i = \sum_{\mathrm{j}} w_{ij} y_j \tag{2}$$

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

Within each layer competition is graded rather than winner-take-all, and is implemented in two stages. First, to implement lateral inhibition the activation 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

$$I_{a,b} = \begin{cases} -\delta \mathrm{e}^{-\frac{a^2+b^2}{\sigma^2}}, & \text{if } a \neq 0 \text{ or } b \neq 0, \\ 1 - \sum\limits_{\substack{a\neq 0 \\ b\neq 0}} I_{a,b}, & \text{if } a = 0 \text{ and } b = 0. \end{cases} \tag{3}$$

The lateral inhibition parameters are given in Table 3.

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

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

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. 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 parameters for the sigmoid activation function are shown in Table 4.

Table 3 Lateral inhibition parameters

Layer	Radius σ	Contrast δ
1	1.38	1.5
2	2.7	1.5
3	4.0	1.6
4	6.0	1.4

Table 4 Sigmoid activation functions

Layer	Percentile	Slope β
1	99.2	190
2	98	40
3	88	75
4	91	26

2.2 Continuous transformation (CT) learning

Continuous transformation (CT) learning utilizes spatial continuity inherent in how objects transform in the real world, combined with associative learning of the feedforward connection weights.

The associative learning is as follows. At each timestep during training, a transform of an object is presented to the retina. The transform may take the form of a shift in location of the image on the retina, or a change in the angle of view of an object, etc. With CT learning, in the early stages of training, the visual stimuli presented to the retina consist of exemplars of each stimulus that reflect small differences in its essentially continuous transformations, for example of position, view, etc. For translation invariance, this means that the retinal images of a stimulus at successive timesteps normally overlap so that two successive images have a number of neurons in the input layer in common. At each timestep, the activity due to the stimulus on the retina is propagated in a feedforward fashion through the network, stimulating patterns of activity in the later layers. Once the activity patterns have been computed in the various layers including competitive lateral inhibition as described above, the synaptic weights of the forward connections between the layers are updated by an associative learning rule which enhances the synaptic weight between two neurons when they are co-firing. There are a variety of associative rules that could be used. In the simulations with CT learning described in this paper we use the 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 normalised at the end of each timestep during training as in usual competitive learning (Hertz et al. 1991).

The CT learning process operates as follows, and is illustrated in Fig. 2. During the presentation of a visual image at one position on the retina that activates neurons in the input layer, a small winning set of neurons in the output layer will modify (through associative learning) their afferent connections from the input layer to respond well to that image in that location. When the same image appears later at nearby locations, so that there is spatial continuity, the same neurons in the output layer will be activated because some of the active afferents are the same as when the image was in the first position. The key point is that if these afferent connections have been strengthened sufficiently while the image is in the first location, then these connections will be able to continue to activate the same neurons in the output layer when the image appears in overlapping nearby locations. The newly active afferents that have just become active because of the transform then show associative synaptic modification onto the same postsynaptic neuron that is active as a result of the part of the stimulus that overlaps with the previous

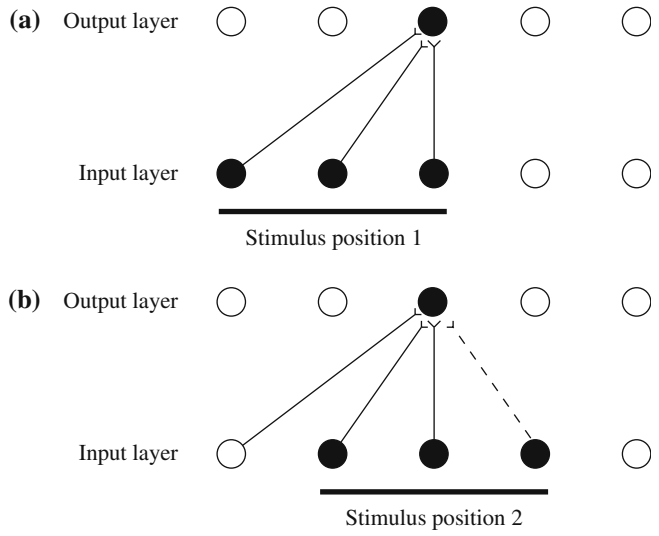

Fig. 2 An illustration of how Continuous transformation (CT) learning would function in a network with a single layer of forward synaptic connections between an input layer of neurons and an output layer. Initially the forward synaptic weights are set to random values. **a** The initial presentation 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** Shows 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

presentation. Thus the same neurons in the output layer have learned to respond to inputs that have similar vector elements in common. We note that the postsynaptic neuron now has strong connections from both the first and the second transform of the stimulus, and the effects of all transforms remain. (The weight normalization referred to above that is useful for competitive learning systems tends to have a weak effect in making synaptic changes that occur early in the learning relatively less strong than those that occur later, and this was not a significant factor in any of the simulations described.) As can be seen in Fig. 2, the process can be continued for subsequent shifts, provided that a sufficient proportion of input cells stay active between individual shifts. This whole process is repeated throughout the network, both horizontally as the image moves on the retina, and hierarchically up through the network. Over a series of stages, transform invariant (e.g. location invariant) representations of images are successfully learned, allowing the network to perform invariant object recognition. A similar CT learning process may operate for other kinds of transformation, such as change in view or size. In this paper we demonstrate CT learning of view invariant representations.

We show in this paper, with supporting simulations, that CT learning can learn large numbers of transforms of objects and does not require any short-term memory trace in the learning rule (Experiments 1 and 2), requires continuity in space but not necessarily in time (Experiment 3), can cope when object transforms are presented in a randomized order (Experiment 4), and can learn objects with just a few exemplar views provided that early layers of the network have been pretrained to provide locally invariant representations of features or feature combinations (Experiment 5).

Once limited invariant responses have been learned by the early layers of the network, CT learning in the higher layers can operate with larger (less continuous) transformations of the stimuli between learning updates. This is because, with invariant responses already learned in the lower layers, a relatively large transformation (e.g. translation) will still activate many of the same neurons in the lower layers due to their transform invariant responses. This means the higher layers will still receive similar inputs before and after the stimulus transform.

2.3 Trace learning

Continuous transform (CT) learning is compared with another approach to invariance learning, trace learning, in this paper, and we summarise next the trace learning procedure developed and analysed previously (Földiák 1991; Rolls 1992; Rolls and Milward 2000; Rolls and Stringer 2001; Wallis and Rolls 1997). Trace learning utilises the temporal continuity of objects in the world (over short time periods) to help the learning of invariant representations. The concept here is that on the short time scale, for e.g. a few seconds, the visual input is more likely to be from different transforms of the same object, rather than from a different object. A theory used to account for the development of view invariant representations in the ventral visual system uses this temporal continuity in a trace learning rule (Rolls and Milward 2000; Rolls and Stringer 2001; Wallis and Rolls 1997). The trace learning mechanism relies on associative learning rules, which utilise a temporal trace of activity in the postsynaptic neuron (Rolls 1992; Földiák 1991). Trace learning encourages neurons to respond to input patterns which occur close together in time, which are likely to represent different transforms (views) of the same object.

The trace learning rule (Rolls 1992; Földiák 1991; Rolls and Milward 2000; Wallis and Rolls 1997) encourages neurons to develop invariant responses to input patterns that tended to occur close together in time, because these are likely to be from the same object. The particular rule used (see Rolls and 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 η was set to 0.8. A discussion of the good performance of this rule, and its relation to other versions of trace learning rules, are provided by Rolls and Milward (2000) and Rolls and Stringer (2001).

The CT learning procedure described above has two major differences from trace learning. Firstly, the visual stimuli presented to the retina must transform continuously, that is there must be considerable similarity in the neurons in layer 2 activated in the competitive process by close exemplars in layer 1. Secondly, in CT learning the synaptic weights are updated by an associative learning rule without a temporal trace of neuronal activity. Thus, without the need for a temporal trace of neuronal activity, different retinal transforms of an object become associated with a single set of invariant cells in the upper layers. We also argue that CT learning can complement trace learning, since trace but not CT learning can associate completely different retinal images that tend to occur close together in time.

2.4 Simulations: stimuli

The stimuli used to train the networks were images of continuously rotating 3D objects. The objects were created using OpenGL, which gives a maximum of control over all stimulus parameters. In this way it was possible to fine-tune the amount by which each stimulus was rotated between views of the objects. OpenGL builds a 3D representation of the objects, and then is able to project different views onto a 2D image. Lighting was mainly ambient with a diffuse light source added to allow different surfaces to be shown with different intensities as illustrated in Fig. 3, which illustrates the two objects used, a cube and a tetrahedron, rotating with a step size of 18°. For all experiments the stimuli were rotated round their vertical axis over a range of 180° (which is sufficient to provide all of the different views of these objects due to their rotational symmetry). In the experiments described here just two stimuli were used, as these are sufficient to demonstrate some of the major properties of CT learning. However, further investigations have already shown that CT learning can operate with larger numbers of objects in the training set.

2.5 Simulations: 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

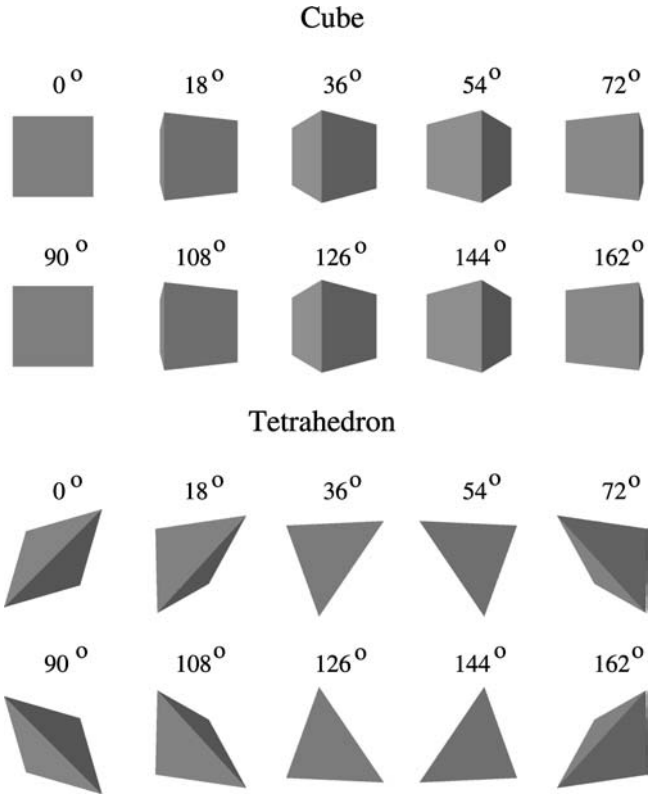

Fig. 3 Views of the two objects used in the simulations, (*top*) a cube, and (*bottom*) a tetrahedron. The network was exposed to views of rotations around the vertical axis with variable step sizes between the views (18° in this figure)

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, 2, 3 and 4 were 50, 100, 100 and 75, respectively. In each epoch all training patterns, i.e. all views of each object, were presented. The learning rates α in Eqs. 6 and 5 for layers 1, 2, 3 and 4 were 0.09, 0.067, 0.05 and 0.04, respectively.

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 (see Rolls and Milward 2000). 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 et al. (1997b) with the following exceptions. First, no correction was made for the limited number of trials because, in VisNet, each measurement of a response is exact, with no variation due to sampling on different trials. Second, the binning procedure was to use equispaced rather than equipopulated bins. This small modification was useful because the data provided by VisNet can produce perfectly discriminating responses with little trial-to-trial variability. Because the cells in VisNet can have bimodally distributed responses, equipopulated bins could fail to perfectly separate the two modes. (This is because one of the equipopulated bins might contain responses from both of the modes.) The number of bins used was equal to or less than the number of trials per stimulus, that is for VisNet the number of positions on the retina (Rolls et al. 1997b). Because VisNet operates as a form of competitive net to perform categorisation of the inputs received, good performance of a neuron will be characterised by large responses to one or a few stimuli regardless of their position on the retina (or other transform), 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, rather than the average amount of information it conveys about the whole set S of stimuli (known as the mutual information). Thus for each cell the performance measure was the maximum amount of information a cell conveyed about any one stimulus (with a check, in practice always satisfied, that the cell had a large response to that stimulus, as a large response is what a correctly operating competitive net should produce to an identified category). In many of the graphs in this paper, the amount of information each of the cells in layer 4 had about any stimulus is shown.

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 all the 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 by Rolls et al. (1997a) and Rolls and Milward (2000). The multiple cell information measure is the mutual information $I(S, \mathbf{R})$, i.e. 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, **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 $P(\mathbf{r}, s)$ embodying the relationship between a stimulus s and the response rate vector $\mathbf{r}$ provided by the firing of the set of neurons to a presentation of that stimulus. (Note that 'stimulus' refers to an individual object that can occur with different transforms, e.g. translation or size, see Wallis & Rolls, 1997). This is 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 $\mathbf{s}'$ that gave rise to the particular firing rate response vector on each trial is estimated. This involves for example maximum likelihood estimation or dot product 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 using a dot product comparison which stimulus was shown. The probabilities of it being each of the stimuli can be estimated in this way. Details are provided by Rolls et al. (1997a). 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

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

In the experiments presented later, the multiple cell information was calculated from only a small subset of the output cells. There were five cells selected for each stimulus, and these were the five cells which gave the highest single cell information values for that stimulus.

We demonstrate the ability of the new CT learning algorithm to train the network to recognise two different 3D stimuli, a cube and tetrahedron, as they are rotated through 180°. The maximum single cell information measure is:

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

where in this case the number of stimuli is 2. This gives a maximum single cell information measure of 1 bit.

3 Results: VisNet simulations

3.1 Experiment 1: demonstration of CT Learning

Experiment 1 provides a demonstration of CT learning. The network was trained with the cube and the tetrahedron using the Hebb rule (5). During training, the cube and the tetrahedron were presented to the network by rotating them continuously round a vertical axis as shown in Fig. 3 in an anticlockwise (from above) direction through 180° in 1° step sizes. This continuous change in viewing angle with small step sizes allows the CT learning effect to work.

Numerical results for Experiment 1 are given in Fig. 4, which shows typical neuron response profiles after training. Figure 4a shows the firing rate response profile of a 4th layer neuron (at coordinates 10,10) to the cube and tetrahedron stimuli as they are rotated through 180°. This neuron has learned to respond to the cube in all viewing angles, and does not respond to the tetrahedron from any view. Therefore, this neuron has learned complete view invariance.

Figure 4b shows the firing rate response profiles of two fourth layer cells which have learned to respond to the tetrahedron. The plot on the left shows the response profile of cell (22,13), which has learned to respond to the tetrahedron over a central region of views covering approximately

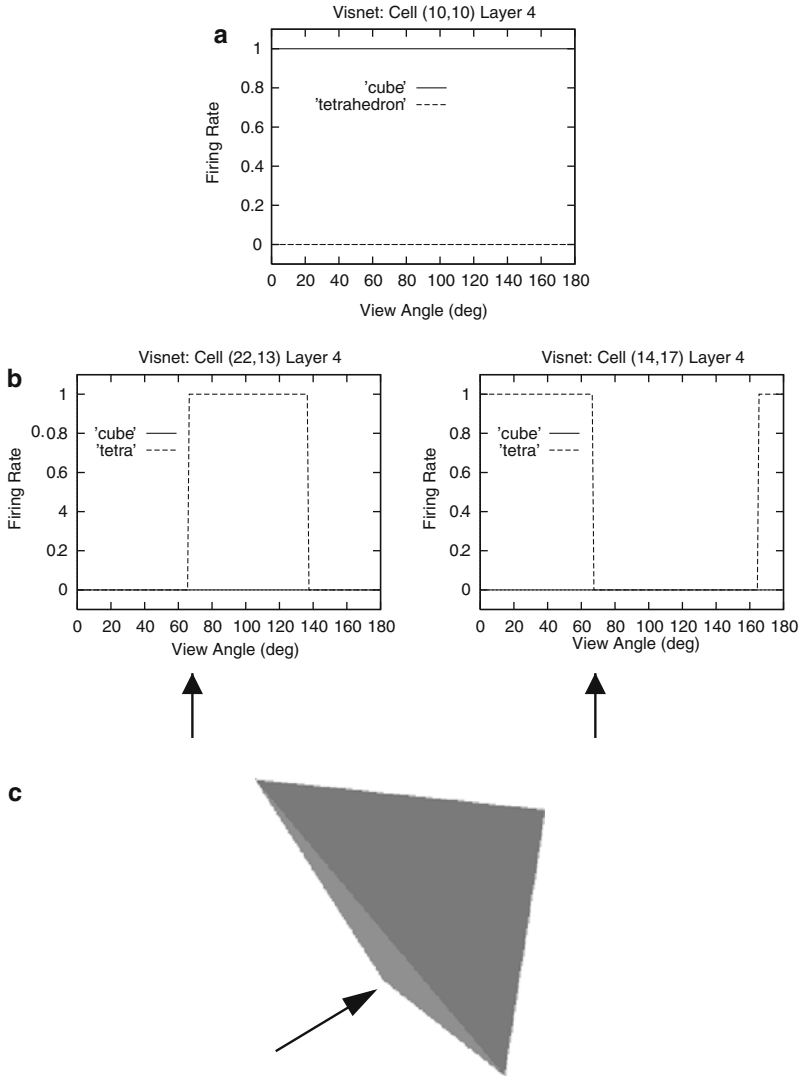

Fig. 4 Experiment 1: Demonstration of CT learning. During training of the network using the Hebb rule Eq. 5, the cube and the tetrahedron are presented to the network by rotating them continuously about a vertical axis in an anticlockwise direction (viewed from above) through 180° in 1° step sizes. **a** The firing rate response profile of a fourth layer neuron (at coordinates 10,10) to the cube and tetrahedron stimuli as they are rotated through 180°. This cell has learned to respond to the cube in all viewing angles, but does not respond to the tetrahedron from any view. **b** The firing rate response profiles of two fourth layer cells that have learned to respond to the tetrahedron. The plot on the left shows the response profile of cell (22,13), which has learned to respond to the tetrahedron over a central region of views covering approximately 65–140°. The plot on the *right* shows the response profile of cell (14,17), which has learned to respond to the tetrahedron over two regions covering approximately 0–65 and 165–180°. **c** The view of the tetrahedron at 65°, with an *arrow* pointing to the new face that has just come into view

65–140°. The plot on the right shows the response profile of cell (14,17), which has learned to respond to the tetrahedron over two extremal regions covering approximately 0–65° and 165–180°, where the view at 180° is identical to the view at 0°.

The responses of these two cells appear to cover disjoint subregions of the viewing space that may be defined by the specific surfaces currently in view. For example, note that cell (14,17) fires from 0 to 65° where it stops firing, while cell (22,13) begins firing at 65° and continues firing until about 140°. The change at 65°, marked by a vertical arrow, occurs when a new face comes into view. Figure 4c shows the view of the tetrahedron at 65°, with an arrow pointing to the new face that has just come into view. Interestingly, this face remains in view until just after the viewing angle reaches 140°, at which point cell (22,13) ceases to fire. For different neurons that responded to the tetrahedron, the discontinuities at which their firing changed tended to be close to (though not typically exactly at) the views at which different surfaces went out of and came into view, i.e. where there were catastrophic changes (see e.g. Koenderink 1990) in the images with view. This is of interest, for it is consistent with the hypothesis that CT learning operates well across continuous or metric (see Biederman 1987) changes in the view

properties, but performs less well across large discontinuities in images.

3.2 Experiment 2: effects of varying the step size between successive viewing angles during training – comparison between CT learning and trace learning

The mechanism of CT learning illustrated in Fig. 2 requires a large overlap in the input patterns which represent nearby transforms of a stimulus. The implication is that CT learning will work with small differences between nearby transforms (e.g. with small changes in view), but will fail when these differences become large. To investigate this quantitatively, we performed Experiment 2 in which we investigated how the change in angle between successive views affected the ability of the network to learn invariant responses to the objects. The performance of networks trained by the Hebb learning rule, and by trace rule learning, was compared. We note that trace rule learning is predicted not to necessarily require similarity between successive inputs, as its learning principle is to associate together stimuli that occur close together in time on the basis that in the real world these stimuli will tend to be different transforms of the same object.

Simulations were performed either using the Hebb rule (5) or using the trace rule (6). During training, the cube and the tetrahedron were presented to the network by rotating them continuously counter-clockwise from above through 180° as illustrated in Fig. 3. For each simulation, the step sizes between successive views were set to one of the following values: 1, 2, 9 and 36°.

Numerical results for Experiment 2 are given in Fig. 5. For Experiment 2 we make use of information measures in order to facilitate the comparison of network performance with the Hebb and trace learning rules. In the left column are the single cell information measures for all top (fourth) layer neurons ranked in order of their invariance to the stimuli. Each row corresponds to a different step size, i.e. 1, 2, 9 and 36°. In the right column are shown the multiple cell information measures. Each plot compares network performance with the Hebb rule (solid line), trace rule (dashed line), and random weights with no learning (dotted line). The simulations with random weights provide a baseline performance with which to compare the performance of the Hebb and trace learning rules.

With a 1° step size (top row of Fig. 5), the single cell information plot shows that the Hebb rule has enabled a number of fourth layer cells to reach a maximal performance of 1 bit. The multiple cell information measures also reach a maximal performance of 1 bit, confirming that an ensemble of layer 4 neurons provide for perfect discrimination between the objects. This is a demonstration of the CT learning mechanism operating for small step sizes between transforms.

Interestingly, the trace rule (6) also performs well with a 1° step size (in Fig. 5 the results for the trace and Hebb rule in the multiple cell information overlap completely). This is because, with a small step size between successive stimulus transforms, the CT effect dominates when using the trace rule. In this case, the trace rule builds view invariance into the network by a form of CT learning. In fact, the trace rule has a Hebbian and a trace component (and is fully Hebbian if $\eta = 0$, with a value of 0.8 used in the simulations). During learning, with successively presented transforms of the same object, 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 $\bar{y}^\tau$ for the active neurons in the higher layers by virtue of Eq. 7. The trace rule (6) can then potentially help to associate the successive views of the same object onto the same active neurons in the higher layers, and in this way potentially contributes usefully to the development of transform invariance in the network. The trace learning rule under conditions in which close transforms of the same object are presented successively can thus help CT learning, for example when the step size is larger, for example 2° in Fig. 5, row 2. If the different transforms of the same object were not presented close in time, then the trace rule is predicted to impair the learning, and this prediction is tested in Experiment 3.

As the differences between transforms get larger, the overlap of activation produced by the closest transforms will become smaller and lead to a breakdown of the CT effect. This is illustrated in Fig. 5, row 2, where with a step size of 2°, the single cell information measures show that the performance with both the Hebb and trace rules is degraded, with no cells reaching the maximum single cell information level of 1 bit. This is because the changes between successive stimulus views are too great to support the CT learning mechanism. The multiple cell information measure in row 2 shows that the trace can encourage neurons to respond to different transforms in this successive presentation paradigm, and can thus produce better performance across a population of neurons than the Hebb rule.

As the step size between views increases from 9 to 36° (rows 3 and 4 of Fig. 5), the performance of the Hebb rule remains poor, that is the CT effect does not work well with these larger differences between closest transforms of the same object. Further, the single and multiple cell information measures reveal that the performance of the trace rule improves and outperforms the Hebb rule for the same reason as with 2° steps shown in row 2. In particular, with a step size of 36°, the trace rule enables a number of cells to reach the maximal single cell information performance of 1 bit. Similarly, the multiple cell information measures also reach the maximum performance of 1 bit. The reason for the superior performance of the trace rule is that with a large step size of 36°, there are only $180/36 = 5$ views of each stimulus for the network to learn invariantly. With only a few views to learn, the trace learning rule (6) is able to support genuine trace learning dynamics, in which neurons in higher layers rely on the temporally traced values of their firing rate in order to learn to respond invariantly to clusters of patterns that tend to occur close together in time, and in this case are different transforms of the same object.

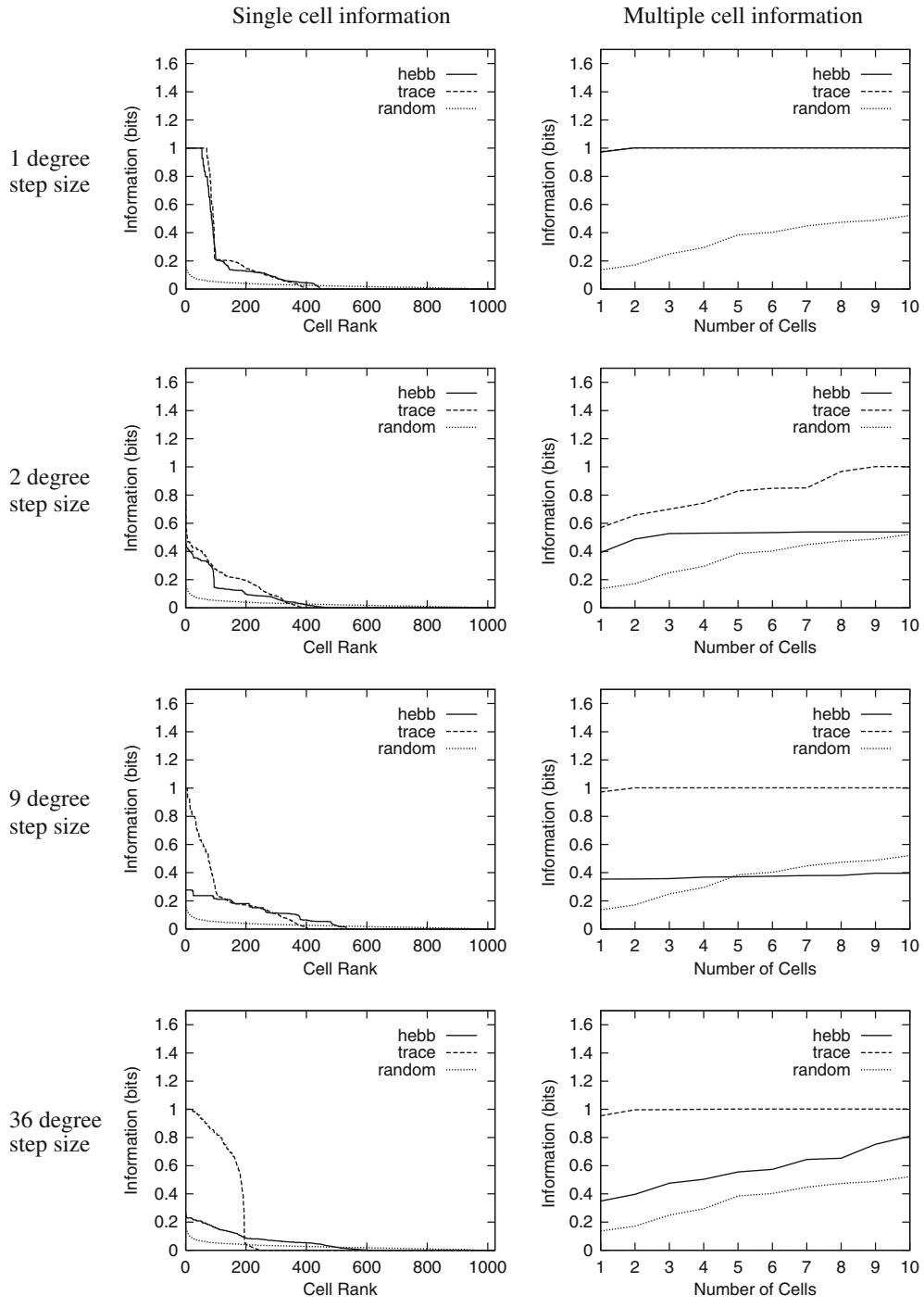

Fig. 5 Experiment 2: Effects of varying the step size between successive viewing angles during training. Simulations were performed either using the Hebb rule Eq 5 or using the trace rule Eq 6. During training, the cube and the tetrahedron were presented to the network by rotating them continuously in an anticlockwise direction through 180°. For each simulation, the step sizes between successive views were set to one of the following values: 1, 2, 9 and 36°. In the *left column* are the single cell information measures for all top (fourth) layer neurons ranked in order of their invariance to the stimuli. In the *right column* are shown the multiple cell information measures. Each *row* corresponds to a different step size, i.e. 1, 2, 9 and 36°. Each plot compares network performance with the Hebb rule (*solid line*), trace rule (*dashed line*), and random weights with no learning (*dotted line*)

The overall picture that emerges from these graphs is that the CT effect can drive performance at small step sizes (e.g. 1° in these simulations) for Hebb rule as well as trace rule learning. With larger rotations between views, the CT effect breaks down as few cells are activated after the competition by the two closest transforms of the same object, so that Hebb

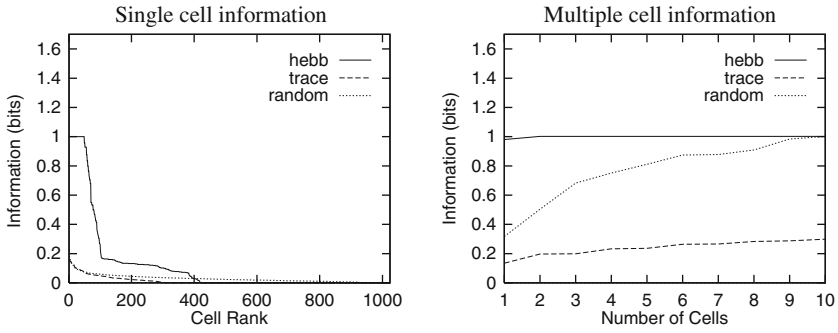

Fig. 6 Experiment 3: Effects of interleaving views of different stimuli during training: comparison between CT learning and trace learning. The network was trained on interleaved views of the cube and tetrahedron over a rotation of 180°, i.e. the first view of the cube was followed by the first view of the tetrahedron, followed by the second view of the cube, followed by the second view of the tetrahedron, and so on. Between views objects were rotated by 1°. On the *left* are the single cell information measures for all top (fourth) layer neurons ranked in order of their invariance to the stimuli. On the *right* are shown the multiple cell information measures. Each plot compares network performance with the Hebb rule (*solid line*), trace rule (*dashed line*), and random weights with no learning (*dotted line*). The single and multiple cell information plots show that the Hebb rule is able to support CT learning, which leads to the development of cells that are able to distinguish the cube and tetrahedron invariant of rotation. In contrast, the trace rule fails to develop invariant cells when the stimuli are interleaved during training

learning cannot contribute to invariant responses. The trace learning rule, on the other hand, improves performance with large step sizes and few views due to genuine trace learning dynamics.

3.3 Experiment 3: effects of interleaving views of different stimuli during training – comparison between CT learning and trace learning

Trace learning relies on temporal continuity between transforms of the same object in the visual environment. CT learning relies on spatial continuity between transforms of the same object in the visual environment. However, with CT learning, one could make the prediction that the spatially closest views need not be presented close together in time, as can be seen from Fig. 2. In Experiment 3 we test the prediction by training with temporally interleaved (i.e. alternated) transforms of different objects. It could be a useful property of the CT learning process that it does not require temporal continuity, as one could imagine a visual scenario in which if there are frequent saccades between different objects, temporal continuity and thus trace learning might be compromised.

In Experiment 3, the cube and the tetrahedron were again rotated over 180° with 1° rotations between individual views, but in contrast to the previous experiments, the different transforms of the two objects were not shown successively for one object and then the other, but were interleaved. (The first view of the cube was followed by the first view of the tetrahedron, followed by the second view of the cube, followed by the second view of the tetrahedron, and so on.)

Numerical results with interleaved training are shown in Fig. 6. The single cell information plot shows that the Hebb rule (5) (solid line) has enabled a number of fourth layer cells to reach a maximal performance of 1 bit. The multiple cell information measures also reach a maximal performance of 1 bit. The Hebb rule is able to support CT learning, which leads to the development of a number of cells that are able to distinguish the cube and tetrahedron invariantly with respect to rotation. Whenever the network is exposed to a new view of a stimulus, even if this is after another object has been shown, there is enough similarity in the input pattern to previously seen views of the first object to be able to activate the same neurons in the higher layers, and this allows these higher layer neurons to learn to respond to the new transform of the first object with the associative Hebb rule. In addition, the difference in features between the two objects seems to be sufficient for the CT effect to dissociate the two objects when shown a set of closely spaced views of each object. That is, the system self-organizes to form a useful representation based just on close similarities of views within an object, learning to associate a new transform of an object with a previously learned transform because the two transforms are sufficiently similar so that the same post-synaptic neuron is activated by both transforms. At the same time, a different object must have sufficiently different views that they are not associated with the views of the first object.

With the trace rule (6) (dashed line), the single and multiple cell information measures are very low. In contrast to the Hebb rule, the trace rule cannot cope with interleaving the stimulus views. This is because the trace rule associates input images which tend to occur close together in time, and thus interleaving the stimulus views would tend to lead to successively shown views of the cube and tetrahedron being associated together.

3.4 Experiment 4: effects of randomising the order in which different views of each stimulus are shown during learning

In the real world, an object might not rotate continuously through all of its transforms each time it is seen. For example, in the real world there might be discontinuous segments of transforms in which the different segments are seen in a random order. If an object was first shown rotating through a 30° segment, and then the same object was shown rotating

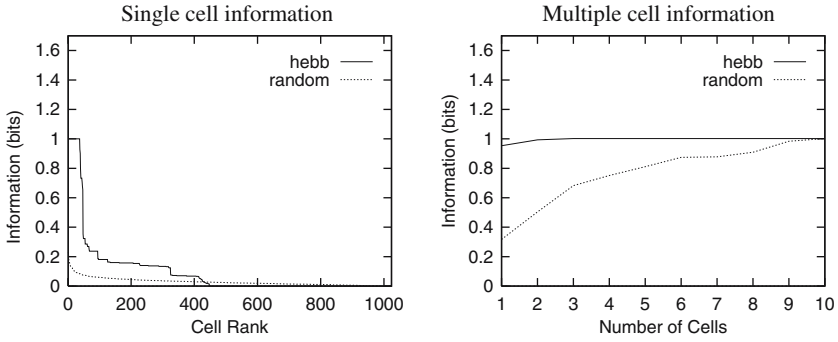

Fig. 7 Experiment 4: Effects of randomising the order in which different views of each stimulus are shown during learning. In this experiment the full set of transformations was divided into six blocks of 30° for each object. Then the network was trained on these blocks presented in random order. The transformations within each block were shown continuously. In the *left graph* are the single cell information measures for all top (fourth) layer neurons ranked in order of their invariance to the stimuli. In the *right graph* are shown the multiple cell information measures. Each plot compares network performance with the Hebb rule (*solid line*), and random weights with no learning (*dotted line*). These results demonstrate that, in principle, CT learning can still occur with a randomised ordering of stimulus views during training

through another 30° segment (chosen from the set of six 30° blocks that are possible for each object), then different neurons might respond to different trial blocks of the same object early on in the training. For example, if 30–60° for object 1 was followed by 0–30° for object 1, then separate neurons in higher layers might respond to the transforms within each of these trial blocks. However, after many training epochs, by chance the 0–30° trial block might be followed by the 30–60° trial block, and this might enable CT learning to make the neurons that respond to trial block 1 learn to respond to trial block 2 of the same object, etc., by remapping the representations for trial block 2 onto the neurons activated by trial block 1. We tested this in Experiment 4. This is an important question, because one might encounter an object from a particular range of viewing angles on one occasion and only later encounter the same object from another range of viewing angles.

For Experiment 4, the views of each object were split up into six blocks. Within each block, the rotation was shown continuously in thirty 1° steps. However, the presentation order of the blocks within and between objects was randomised. As in previous simulations, the number of training epochs for the different layers were 50, 100, 100 and 75, respectively.

The results are shown in Fig. 7, where the learning uses the Hebb rule. The single cell information plot shows that the Hebb rule (5) has enabled a number of fourth layer cells to reach a maximal performance of 1 bit. The multiple cell information measures also reach a maximal performance of 1 bit. These results confirm that the Hebb rule can support CT learning and the development of transform invariance in the network, even when the presentation order of the stimulus views is randomised within blocks. In the simulations for Experiment 4 the learning rates were retuned to 0.0004, 0.001, 0.001, 0.001 in layers 1, 2, 3 and 4, respectively, to improve performance.

These results demonstrate that, in principle, CT learning can still occur with a randomised block ordering of stimulus views during training. This process works by remapping the output representations for different trial blocks, until all of the trial blocks for an object settle on a single output representation over a number of training epochs.

3.5 Experiment 5: neurons in the lower layers, which develop invariant responses to simple features during early visual experience, can help higher layers of the network to generalise to novel stimulus views

In the above experiments, the Hebb rule supported CT learning only when the 180° view space was covered by many closely spaced views during training. For example, in Experiment 2, the Hebb rule developed transform invariant cells when the step size was only 1°, but failed to produce transform invariance when the step size was increased to 2° or larger. This reliance on a large number of closely spaced transforms during training in order to build invariance is a major limitation. The primate visual system, however, can learn to recognise objects invariantly from only a few canonical views.

In Experiment 5 we tested the hypothesis that the later layers of the network need only be trained on a small number of quite different canonical views, as long as the early layers have been pretrained during early visual experience to develop a limited amount of invariance to low-level stimulus features. If the neurons in the early layers have learned to respond invariantly to parts of objects with some limited transform (view) invariance, then the later layers of the network should be able to generalise to novel stimulus views after training on a limited set of canonical stimulus transforms. This has been shown to be possible with trace learning (Stringer and Rolls 2002), and is tested now for CT learning.

In Experiment 5, the first two layers were trained on a full set of 180 transforms with 1° rotations between transforms. Next, in a separate training session, layers 3 and 4 were trained on five non-overlapping canonical views with a step size of 36° rotation between transforms. After these two phases of training, the network was then tested on a full set of stimulus transforms with a step size of 1°, i.e. the stimuli were presented during testing at all 180 views.

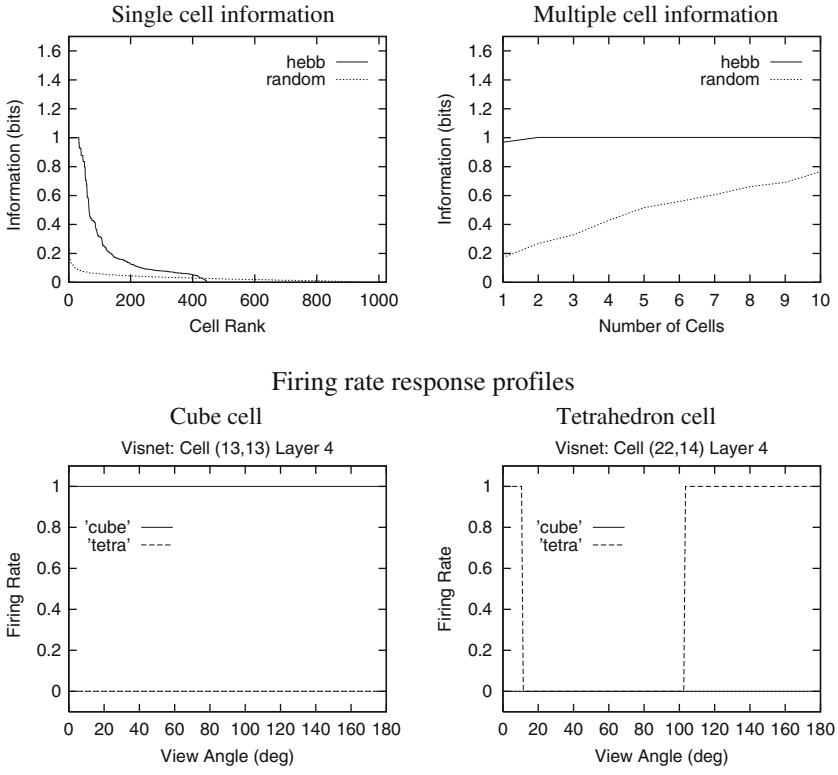

Fig. 8 Experiment 5: Neurons in the lower layers, which develop invariant responses to simple features during early visual experience, can help higher layers of the network to generalise to novel stimulus views. In this simulation the first two layers were trained on a full set of 180 transforms with 1° rotations between transforms. Next, in a separate training session, layers 3 and 4 were trained on five canonical views with a step size of 36° rotation between transforms. After these two phases of training, the network was then tested on a full set of stimulus transforms with a step size of 1°, i.e. the stimuli were presented during testing at all 180 views. *Top row* the single and multiple cell information plots. Each plot compares network performance with the Hebb rule (*solid line*), and random weights with no learning (*dotted line*). Even though layers 3 and 4 had not been exposed to 97% of the views they were tested on, the network achieves good invariance in both single cell and multiple cell information plots as compared to the untrained network condition (*dotted line*). *Bottom row* the firing rates of two fourth layer neurons as a function of the 180 test views. Even though layers 3 and 4 had been trained on only five views, these neurons generalised to many different views of the objects to which they responded, which were the cube for the neuron on the left, and the tetrahedron for the neuron on the right

Numerical results for neurons in the output (fourth) layer of the network are shown in Fig. 8. Even though layers 3 and 4 had not been exposed to 97% of the views they were tested on, the network achieves good invariance in both single cell and multiple cell information plots as compared to the untrained network condition (dotted line). The performance in fact is as good as when training is performed on all 180 locations for layers 1–4, as is illustrated in the top row of Fig. 5. This result is further illustrated in the lower part of Fig. 8 by the firing rates of two single cells in layer 4 to the 180 test views after layers 3 and 4 had been trained only on five views at 0, 36, 72, 108, and 144°. Perfect generalisation to all views is shown for the neuron illustrated which responded to a cube, even though the training of layers 3 and 4 had been at only five views. Generalisation to different views for the neuron illustrated in Fig. 8 which responded to the tetrahedron was as good as that shown in Fig. 4 for Experiment 1 when training of all layers was for all 180 views. These results can be attributed to the fact that training the first two layers on a full set of stimuli has enabled the network to create neurons in layer 2 that generalise within part of the rotation space of an object, leaving layers 3 and 4 for neurons to associate together the different subregions of the rotation space, and also perhaps feature combinations derived from earlier layers.

In order to confirm that training of layers 3 and 4 on the five canonical views was indeed responsible for developing neurons in the fourth layer with invariance to the full set of 180 views with a step size of 1°, we performed the following comparison. We determined whether the invariance developed in layer 2 during initial training of layers 1 and 2 on the full set of 180 views was less than that found in layer 4 after training layers 3 and 4 on five views (which is illustrated in Fig. 8). We were able to confirm that the view invariance for cells in the second layer shown in Fig. 9 was much less than that in layer 4. The single cell information measures for layer 2 are not as good as those for layer 4. These results confirm that the training of layers 3 and 4 on the five canonical views was indeed responsible for developing many neurons in the fourth layer with invariance to the full set of 180 views.

4 Discussion

Continuous transformation (CT) learning is a new algorithm for unsupervised training, which relies on continual synap-

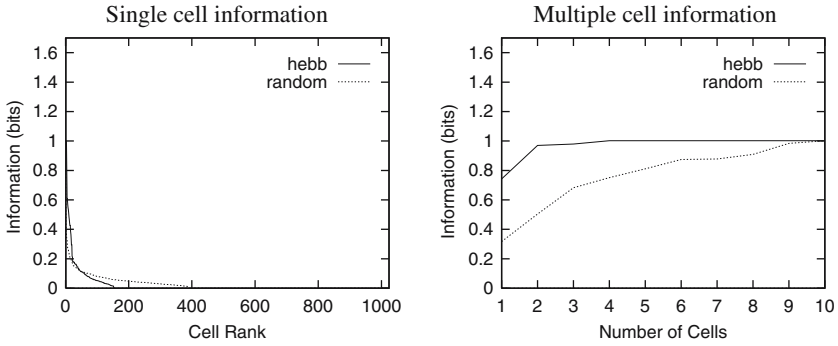

Fig. 9 Experiment 5: Neurons in the lower layers, which develop invariant responses to simple features during early visual experience, can help higher layers of the network to generalize to novel stimulus views. Information measures for the neurons in the second layer of the network are shown. The neurons in the second layer have been trained and tested on the full set of stimulus transforms with a step size of 1°, i.e. the stimuli were presented during testing at all 180 views. In the *left graph* are the single cell information measures for all second layer neurons ranked in order of their invariance to the stimuli. In the *right graph* are shown the multiple cell information measures for the second layer cells. Each plot compares network performance with the Hebb rule (*solid line*), and random weights with no learning (*dotted line*). The single cell information measures are very low

tic modification of the feedforward inter-layer connection weights using an associative (e.g. Hebbian) learning rule during continuous transformation (e.g. translation, rotation, etc.) of the visual stimulus. The aspect in which continuity is implied by CT learning is that adjacent transforms in the space of transforms must be sufficiently close so that the same post-synaptic neuron in the next layer is activated by both transforms after learning to at least one of the transforms has occurred. The condition for this to occur is that the overlap between the two input vectors for two close transforms must be high, so that after the post-synaptic neuron has been made to fire by the first transform, and Hebbian learning has occurred to enhance the synapses from all the active inputs on to that neuron, then the second transform will activate the same output neuron through its shared strengthened synapses with the first transform. At the same time, the overlap with the closest exemplar of a different object must be sufficiently low so that the second object does not produce firing in the same second layer neuron.

Once limited invariant responses have been learned by the early layers of the network, CT learning in the higher layers can operate with larger (less continuous) transformations of the stimuli between learning updates. This is because, with invariant responses already learned in the lower layers, a relatively large transformation (e.g. translation) will still activate many of the same neurons in the lower layers due to their transform invariant responses. This means the higher layers will still receive similar inputs before and after the stimulus transform.

Continuous transformation (CT) learning is biologically plausible in that it requires a standard Hebb associative learning rule, coupled with a process such as heterosynaptic long-term synaptic depression which has the effect of tending to normalise the synaptic weights on each neuron, which is necessary in competitive learning systems (Hertz et al. 1991; Rolls and Deco 2002). Further, CT learning can in principle learn an invariant representation from a single training epoch (i.e. one presentation of each view of each object).

It is interesting to compare invariant learning with the CT and trace learning mechanisms. CT learning operates well with small steps between the input stimuli (see Fig. 5, top row). Trace rule learning also operates well with this small step size, and this can be attributed mainly to the CT effect (see Fig. 5). As the step size is increased, the invariance becomes less good for both the Hebb and trace rules (see Fig. 5). The reason for this deterioration in performance for the CT effect is that the closest transforms become too distant in terms of their overlap to activate the same higher layer neuron. The reason for the poor performance with step sizes of 2° for the trace learning effect is that there are too many transforms to be associated together by a temporal trace (and CT effects do not operate at this scale). Interestingly, as the transforms become even more different (e.g. 9 and 36° in Fig. 5), the trace rule performance increases, as it can associate together even quite different views as long as they are temporally associated, and as long as there are not too many transforms to be associated together. In a sense, a combination of the two learning processes would be useful. If for each object there is a set of closely spaced transforms, CT learning can provide usefully invariant representations for these. On the other hand, if these spatially similar ranges of views are separated by major discontinuities, such as occur with catastrophically different views of 3D objects as new faces come into view (Koenderink 1990) (such as the inside of a jug or the other side of a card), then trace learning can associate together the different catastrophically different views. It would be very interesting to explore this issue further.

Continuous transformation (CT) learning does not require the stimulus transforms to occur in temporal succession (as shown in Experiment 3 in which interleaved views of different stimuli were used during training). In fact, even if visual fixation moves rapidly and randomly between different views of different objects or faces by saccades, CT learning will still develop invariant representations of the individual stimuli. Trace learning will not learn invariance if under some circumstances different objects are seen in close temporal

proximity as frequently as different transforms of the same object.

An interesting aspect of CT learning is that it greatly increases the number of transforms of a given object that can be learned. This is shown by the better performance with 1° than with 2° steps in Fig. 5. CT learning is of course implied with the trace learning rule with these small step sizes. The implication is that applications with the trace learning rule might benefit from small step sizes. At the same time, the trace learning rule would be preferred to the Hebb rule under some circumstances, because the trace rule can cope with catastrophic view changes. Thus training with the trace rule, and small step sizes, may be an interesting area for further capacity investigations, provided of course that the system is shown different transforms of each object with some temporal continuity present within an object. Trace rule learning can of course operate well if there are just short sequences of different views of each object interleaved with short sequences of a different object, for then the associations within an object are stronger than associations between objects.

With trace learning we have only been able to train the network at a limited number of retinal locations before performance degrades (Wallis and Rolls 1997). However, with CT learning, good performance is maintained over a large training set of exemplar views as shown in this paper. Although we have used two objects, each with many views, in the simulations described here, so that the principles of operation can be kept clear, we have shown already in further simulations that the number of objects can be increased, and indeed have good performance with five objects, each of which is more complicated than the objects used in this first paper to demonstrate the principles of continuous transformation learning. These further simulations will be the subject of a future paper in which a series of investigations of the capacity of the system when trained with continuous transformation learning will be presented.

The CT learning does not even require objects to transform smoothly in space across continuous time, as long as all intermediate views are eventually seen in some random order. In fact, it is not even necessary to see all the transforms that might smoothly and continuously cover a space, as long as the early/intermediate layers of the network have already developed some low level feature invariance (as shown in Experiment 5). In this case, previously unseen views of previously trained objects can still be recognised by the network (Stringer and Rolls 2002).

Furthermore, because CT learning does not require a trace $\bar{y}$, there is no parameter η that must be tuned to match presentation sequence length. Instead, the CT object learning will begin automatically at the presentation of a new object, and continue learning invariance to that object until there is a sudden large change in the representation in the lower layers signifying a new object. However, this process will only work if successive representations of the object in the lower layers have many neurons in common.

The results of Experiment 4 demonstrate that, in principle, CT learning can still occur with a randomised block ordering of stimulus views during training. However, the statistics of how stimulus presentations are randomised may well be critical to whether or not the network is able to develop invariance during training. Much more work is needed to characterise what kinds of presentation statistics will allow invariance to develop.

We note that CT learning as described here requires no trace (which could be explicit in the learning rule, or implicit in the continuing firing of the post-synaptic neurons between successive transforms of the same object), for learning. In VisNet the form of the trace can be implemented in a number of ways (Rolls and Milward 2000; Rolls and Stringer 2001; Wallis and Rolls 1997), and use of a trace has been adopted by others (Bartlett and Sejnowski 1998; Becker 1999) and is implicit in the "persistence of unit activity from one cycle to the next" in the network studied by Almassy et al. (1998).

There are interesting spatial constraints that underlie continuous transformation learning. As is evident from Fig. 2, each input neuron must respond (to whatever its effective feature is) when the input stimulus is transformed by at least one position. It is this that helps neurons in the output layer to retain their firing to the transformed version of the stimulus moved by one position, and thereby to benefit from spatial continuity to learn invariant representations. The implication is that the receptive fields of the input layer neurons need to be wider than the spatial offset of the centers of the receptive fields from each other. This is consistent with the receptive field sizes of V1 neurons, the cortical magnification factor (Rolls and Cowey 1970), and the density of neurons in the cortex (Rolls and Deco 2002).

We note that in the VisNet architecture, the invariant representations arise at each layer as a result of either the temporal continuity (learned by the trace rule) or the spatial continuity (learned by continuous transformation learning) over whatever information converges from the previous layer. In contrast, other systems use different methods to wire-in the invariance, such as complex cells in each layer that respond to a sum of offset simple cell inputs (Fukushima and Tanigawa 1996; Fukushima 2003).

It is an important aspect of the architecture described that the representations that are learned are kept sparse (by for example competition), so that every stimulus does not become associated with every other stimulus. The system must learn to represent the rare statistical regularities in the high dimensional space of possible patterns.

Given that there are topological maps in early cortical (e.g. visual) areas with locally convergent feedforward connectivity, and that associative learning (which is reflected in long-term potentiation) is common in the cerebral cortex, (Artola and Singer 1993; Singer 1995; Frégnac 1996), a process with the general characteristics of CT learning is implied to be a quite general property of the functional architecture of the cerebral cortex, provided that the input stimuli transform in small steps. Although CT learning is thus very likely to occur with locally convergent feedforward connectivity, we do note that CT learning can nevertheless occur in non-topologically mapped systems, in which the connectivity allows

the same postsynaptic neuron to receive inputs from neurons that could be anywhere in the input representation, but would be likely to be activated in a way that reflects even quite abstract continuity in the input space.

Acknowledgements This research was supported by the Wellcome Trust, and by the MRC Interdisciplinary Research Centre for Cognitive Neuroscience. GP is an MRC-supported graduate student.

References

Almassy N, Edelman GM, Sporns O (1998) Behavioral constraints in the development of neuronal properties: a cortical model embedded in a real-world device. Cereb Cortex 8:346–361

Artola A, Singer W (1993) Long-term depression of excitatory synaptic transmission and its relationship to long-term potentiation. Trends Neurosci 16:480–487

Bartlett MS, Sejnowski TJ (1998) Learning viewpoint-invariant face representations from visual experience in an attractor network. Netw Comput Neural Syst 9:399–417

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

Biederman I (1987) Recognition-by-components: a theory of human image understanding. Psychol Rev 94:115–147

Booth MCA, Rolls ET (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. J Cognit Neurosci 3:1–8

Elliffe MCM, Rolls ET, Stringer SM (2002) Invariant recognition of feature combinations in the visual system. Biol Cybern 86:59–71

Földiák P (1991) Learning invariance from transformation sequences. Neural Comput 3:194–200

Frégnac Y (1996) Dynamics of cortical connectivity in visual cortical networks: an overview. J Physiol Paris 90:113–139

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

Fukushima K (2003) Neocognitron for handwritten digit recognition. Neurocomputing 51:161–180

Fukushima K, Tanigawa M (1996) Use of different thresholds in learning and recognition. Neurocomputing 11:1–17

Hasselmo ME, Rolls ET, Baylis GC, Nalwa V (1989) Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. Exp Brain Res 75:417–429

Hertz J, Krogh A, Palmer RG (1991) Introduction to the theory of neural computation. Addison Wesley, Wokingham, UK

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

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

Koenderink JJ (1990) Solid shape. MIT, Cambridge

Op de Beeck H, Vogels R (2000) Spatial sensitivity of macaque inferior temporal neurons. J Comp Neurol 426:505–518

Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2:1019–1025

Rolls ET (1992) Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. Phil Trans Roy Soc 335:11–21

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

Rolls ET (2005) Emotion explained. Oxford University Press, Oxford

Rolls ET, Baylis GC (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. Exp Brain Res 65:38–48

Rolls ET, Cowey A (1970) Topography of the retina and striate cortex and its relationship to visual acuity in rhesus monkeys and squirrel monkeys. Exp Brain Res 10:298–310

Rolls ET, Baylis GC, Hasselmo ME (1987) The responses of neurons in the cortex in the superior temporal sulcus of the monkey to band-pass spatial frequency filtered faces. Vis Res 27:311–326

Rolls ET, Baylis GC, Leonard CM (1985) Role of low and high spatial frequencies in the face-selective responses of neurons in the cortex in the superior temporal sulcus. Vis Res 25:1021–1035

Rolls ET, Deco G (2002) Computational neuroscience of vision. Oxford University Press, Oxford

Rolls ET, 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 Comput 12:2547–2572

Rolls ET, Stringer SM (2001) Invariant object recognition in the visual system with error correction and temporal difference learning. Netw Comput Neural Syst 12:111–129

Rolls ET, Treves A, Tovee MJ (1997a) The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. Exp Brain Res 114:149–162

Rolls ET, Treves A, Tovee M, Panzeri S (1997b) Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. J Comput Neurosci 4:309–333

Singer W (1995) Development and plasticity of cortical processing architectures. Science 270:758–764

Stringer SM, Rolls ET (2002) Invariant object recognition in the visual system with novel views of 3D objects. Neural Comput 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. J Neurophysiol 66:170–189

Tovee MJ, Rolls ET, Azzopardi P (1994) Translation invariance and the responses of neurons in the temporal visual cortical areas of primates. J Neurophysiol 72:1049–1060

Ullman S (1996) High-level vision. MIT, Cambridge

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

Wallis G, Rolls ET (1997) Invariant face and object recognition in the visual system. Progr Neurobiol 51:167–194