

# Continuous transformation learning of translation invariant representations

G. Perry · E. T. Rolls · S. M. Stringer

Received: 4 February 2009 / Accepted: 21 May 2010 / Published online: 11 June 2010  
© Springer-Verlag 2010

**Abstract** We show that spatial continuity can enable a network to learn translation invariant representations of objects by self-organization in a hierarchical model of cortical processing in the ventral visual system. During ‘continuous transformation learning’, the active synapses from each overlapping transform are associatively modified onto the set of postsynaptic neurons. Because other transforms of the same object 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 show that the transforms must be close for this to occur; that the temporal order of presentation of each transformed image during training is not crucial for learning to occur; that relatively large numbers of transforms can be learned; and that such continuous transformation learning can be usefully combined with temporal trace training.

**Keywords** Object recognition · Continuous transformation · Trace learning · Inferior temporal cortex · Invariant representations

## Introduction

Over successive stages, the visual system develops neurons that respond with position (i.e. translation), view, and size invariance to objects or faces (Desimone 1991; Rolls 1992, 2000, 2007, 2008a, b, 2010; Tanaka et al. 1991; Rolls and Deco 2002). For example, it has been shown that the inferior temporal visual cortex has neurons that respond to faces and objects invariantly with respect to translation (Tovee et al. 1994; Kobatake and Tanaka 1994; Ito et al. 1995; Op de Beeck and Vogels 2000; Rolls et al. 2003), 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). It is crucially important that the visual system builds invariant representations, for only then can associative learning on one trial about an object generalize usefully to other transforms of the same object (Rolls and Deco 2002; Rolls 2008b). 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 (Rolls and Deco 2002; Ullman 1996; Riesenhuber and Poggio 1999b; Biederman 1987; Bartlett and Sejnowski 1998; Becker 1999; Wiskott and Sejnowski 2002; Rolls et al. 2008; Rolls 2009).

In this paper, we show how a quite general learning principle, *Continuous Transformation (CT) learning* (Stringer et al. 2006), could be used to build translation invariant representations. Continuous spatial transformation learning utilizes spatial continuity of objects, in the world, in contrast to previous approaches which have used temporal continuity, for example, using a modified Hebb rule with a short term temporal trace of preceding activity (Földiák 1991; Wallis and Rolls 1997; Rolls and Milward

---

G. Perry · E. T. Rolls (✉) · S. M. Stringer  
Centre for Computational Neuroscience,  
Department of Experimental Psychology,  
Oxford University, South Parks Road,  
Oxford OX1 3UD, England  
e-mail: Edmund.Rolls@oxcns.org  
URL: <http://www.oxcns.org>

2000; Rolls and Stringer 2001; Rolls and Deco 2002). The CT-based learning we describe here can be powerful, for it relies on spatial overlap between stimuli in small regions of the input space, but given training exemplars throughout the space enables transforms in quite distant parts of the continuous input space to be associated together onto the same population of postsynaptic neurons. We show how continuous transformation learning could be used in the type of hierarchical processing that is a property of cortical architecture, in which key principles agreed by many investigators (Fukushima 1980; Wallis and Rolls 1997; Riesenhuber and Poggio 1999a, b, 2000; Giese and Poggio 2003; Serre et al. 2007) include feedforward connectivity, local lateral inhibition within a layer to implement competition, and then some form of associative learning. We show by simulation how CT learning can be used to build translation invariant representations in a hierarchical network model (VisNet) of cortical processing in the ventral visual system, and we show how CT learning can usefully complement trace rule learning (Földiák 1991; Wallis and Rolls 1997; Rolls and Milward 2000; Rolls and Stringer 2001; Rolls and Deco 2002). We note that in this and previous papers with VisNet, we show that invariant object recognition learning can occur in purely feedforward hierarchical networks, but we have extended elsewhere the model to include top-down backprojections to model top-down attentional effects in visual object recognition (Deco and Rolls 2004, 2005; Rolls 2008b), and have postulated elsewhere that top-down backprojections could be useful to guide learning in hierarchical networks (Rolls and Treves 1998; Rolls and Deco 2002).

In previous papers on continuous transformation learning, we investigated view invariance (Stringer et al. 2006; Perry et al. 2006). In this paper, we address for the first time with continuous transformation learning a type of invariance learning that is fundamental to visual object recognition, namely translation invariance learning. Translation invariance in two dimensions is a hard problem to solve because any individual feature is likely to occur in every possible location due to the shifts in position, and this leads to great overlap at the feature level between the different transforms of different objects. Hence, we consider translation invariance in two dimensions to be a stronger challenge for continuous transformation learning. In addition, it is particularly important to test CT learning on the problem of translation invariance learning, because the changes with each transform can be clearly defined at the pixel level, whereas with view transforms, the changes at the pixel level between adjacent transforms are less clearly specified as an object is rotated. Translation invariance thus offers a way to assess better how far apart the transforms of objects can be for CT learning to still operate effectively and how many different transforms of a given object can be learned by CT

learning. Another novel aspect of the work described here is that we also describe conceptually how invariance learning requires different weight vectors to be learned by neurons than are typically learned by competitive networks in which the patterns within a cluster overlap with each other. We show here how translation invariant representations can be learned in continuous transformation learning by the associative processes used in a standard competitive network and where there may be no overlap between the patterns at the extremes of the transform of a given object. Another new aspect of the work described here is that we show how continuous transformation learning can usefully complement trace rule learning to form invariant representations of objects.

The aim of the research described here is thus to investigate whether spatial continuity in the world can provide a basis for the formation of translation invariant representations that discriminate between objects. The aim is not to investigate how many objects can be discriminated, for we agree with Pinto et al. (2008) that a more fundamental issue in understanding visual object recognition is to understand how invariant representations can be formed using well-controlled stimuli, rather than large numbers of stimuli. In this paper, we show for the first time that translation invariant object recognition can be formed by a self-organizing competitive process utilizing purely spatial continuity as objects transform.

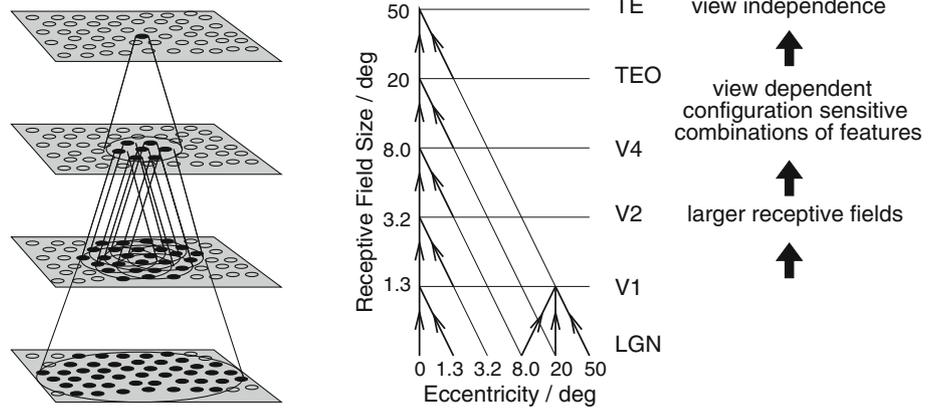
## Methods

### The VisNet architecture

The model architecture (VisNet), implemented by Wallis and Rolls (1997) and Rolls and Milward (2000), that is used to investigate the properties of CT learning in this paper is 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; and (iii) Synaptic plasticity based on a Hebb-like learning rule.

The model consists of a hierarchical series of four layers of competitive networks, intended to model in principle the hierarchy of processing areas in the ventral visual stream which include V2, V4, the posterior inferior temporal cortex, and the anterior inferior temporal cortex, 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

**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 *VI* visual cortex area V1, *TEO* posterior inferior temporal cortex, *TE* inferior temporal cortex (IT)



**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	–	–

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. The input filters used are computed by weighting the difference of two Gaussians by a third orthogonal Gaussian according to the following:

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

where  $f$  is the filter spatial frequency,  $\theta$  is the filter orientation, and  $\rho$  is the sign of the filter, i.e.  $\pm 1$ . Individual filters are tuned to spatial frequency (0.0625–0.5 cycles/pixel); orientation ( $0^\circ$ – $135^\circ$  in steps of  $45^\circ$ ); and sign ( $\pm 1$ ). Filters are thresholded to provide positive only firing rates. Oriented difference of Gaussian filters were chosen in preference to Gabor filters on the grounds of their better fit to available neurophysiological data including the zero DC response (Hawken and Parker 1987). Simple cell like response properties were chosen rather than complex cell like properties because it is important in feature hierarchy networks to build neurons early in processing that respond to combinations of features in the correct relative spatial positions in order to provide a solution to the binding problem (Elliffe et al. 2002; Rolls and Deco 2002). The

**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

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,

$$h_i = \sum_j w_{ij} y_j \tag{2}$$

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 is convolved with a spatial filter,  $I$ , where  $\delta$  controls the contrast and  $\sigma$  controls the width, and  $a$  and  $b$  index the distance away from the center of the filter

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

The lateral inhibition parameters are given in Table 3.

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

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

where  $r$  is the activation (or firing rate) after lateral inhibition,  $y$  is the firing rate after contrast enhancement, and  $\alpha$  and  $\beta$  are the sigmoid threshold and slope, respectively. The parameters  $\alpha$  and  $\beta$  are constant within each layer, although  $\alpha$  is adjusted at each iteration 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

**Table 3** Lateral inhibition parameters

Layer	1	2	3	4
Radius, $\sigma$	1.38	2.7	4.0	6.0
Contrast, $\delta$	1.5	1.5	1.6	1.4

of the activations within the layer. The parameters for the sigmoid activation function are shown in Table 4.

Model simulations that 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 (Wallis and Rolls 1997; Rolls and Milward 2000; Rolls and Stringer 2001).

In this paper, the CT learning principle implemented 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.

### Continuous transformation 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 network architecture is that of a competitive network (Hertz et al. 1991; Rolls and Deco 2002).

The continuous transformation learning process is illustrated for translation invariance learning in Fig. 2. During the presentation of a visual image at one position on the retina that activates neurons in layer 1, a small winning set of neurons in layer 2 will modify (through associative learning) their afferent connections from layer 1 to respond well to that image in that location. A variety of associative rules could be used. In the simulations with CT learning described in this paper, we use the Hebb learning rule

$$\delta w_{ij} = \alpha y_i x_j \quad (5)$$

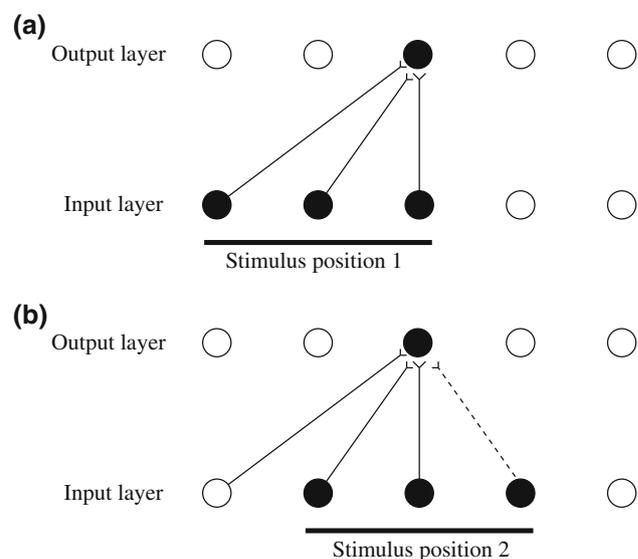
where  $\delta w_{ij}$  is the increment in the synaptic weight  $w_{ij}$ ,  $y_i$  is the firing rate of the postsynaptic neuron  $i$ ,  $x_j$  is the firing rate of the pre-synaptic neuron  $j$ , and  $\alpha$  is the learning rate. To bound the growth of each neuron's synaptic weight

**Table 4** Sigmoid parameters

Layer	1	2	3	4
Percentile	99.2	98	88	91
Slope $\beta$	190	40	75	26

vector,  $w_i$  for the  $i$ th neuron, its length is normalized at the end of each timestep during training as in standard competitive learning (Hertz et al. 1991). When the same image appears later at nearby locations, so that there is spatial continuity, the same neurons in layer 2 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 layer 2 when the image appears in overlapping locations. Thus, the same neurons in the output layer have learned to respond to inputs that have similar vector elements in common.

As can be seen in Fig. 2, the process can be continued for subsequent shifts, provided that a sufficient proportion



**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. The *top part a* shows 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. The *bottom part 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

of presynaptic neurons stay active between individual shifts. This learning process in a hierarchical network can take place at every level of the network (Wallis and Rolls 1997; Rolls and Deco 2002; Rolls and Stringer 2006; Rolls 2008b). 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.

This paper describes the first investigation of CT learning of translation invariant representations. We show, with supporting simulations, that CT learning can implement translation invariance learning of objects, can learn large numbers of transforms of objects, does not require any short term memory trace in the learning rule, requires continuity in space but not necessarily in time, and can cope when the transforms of an object are presented in a randomized order.

### Trace learning

CT learning is compared in Experiment 2 with another approach to invariance learning, trace learning, and we summarize next the trace learning procedure developed and analyzed previously (Földiák 1991; Rolls 1992; Wallis and Rolls 1997; Rolls and Milward 2000; Rolls and Stringer 2001). Trace learning utilizes 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, of 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* (Wallis and Rolls 1997; Rolls and Milward 2000; Rolls and Stringer 2001). The trace learning mechanism relies on associative learning rules, which utilize a temporal trace of activity in the postsynaptic neuron (Földiák 1991; Rolls 1992). Trace learning encourages neurons to respond to input patterns which occur close together in time, which are likely to represent different transforms (positions) of the same object. Temporal continuity has also been used in other approaches to invariance learning (Stone 1996; Bartlett and Sejnowski 1998; Becker 1999; Einhäuser et al. 2002; Wiskott and Sejnowski 2002).

The trace learning rule (Földiák 1991; Rolls 1992; Wallis and Rolls 1997; Rolls and Milward 2000) 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

$$\delta w_j = \alpha \bar{y}^{\tau-1} x_j^{\tau} \quad (6)$$

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

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

and we have the following definitions

$x_j$ :  $j$ th input to the neuron.

$\bar{y}^{\tau}$ : Trace value of the output of the neuron at time step  $\tau$ .

$w_j$ : Synaptic weight between  $j$ th input and the neuron.

$y$ : Output from the neuron.

$\alpha$ : Learning rate. Annealed to zero.

$\eta$ : Trace value. The optimal value varies with presentation sequence length.

The parameter  $\eta$  may be set anywhere in the interval [0,1], and for the simulations described here,  $\eta$  was set to 0.8. A discussion of the good performance of this rule, principles by which it can be set to optimal values, and its relation to other versions of trace learning rules are provided by Rolls and Milward (2000), Rolls and Stringer (2001), and Wallis and Baddeley (1997).

The CT learning procedure described previously has two major differences from trace learning. First, 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, as trace but not CT learning can associate completely different retinal images that tend to occur close together in time.

### Invariance learning vs conventional competitive learning

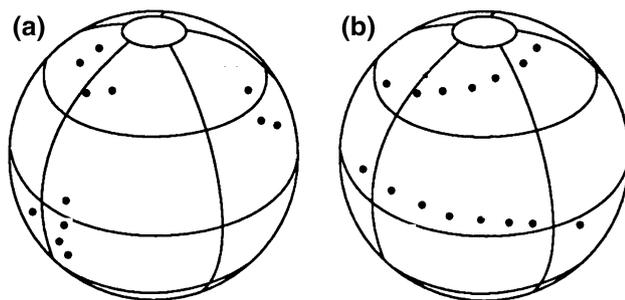
Before presenting results from this study, it is first useful to illustrate a fundamental difference in the weight structure required to solve an invariance task and that required to solve the kind of categorization task that competitive networks have more commonly been used for.

In competitive learning as typically applied, the weight vector of a neuron can be thought of as moving toward the center of a cluster of similar overlapping input stimuli (Rumelhart and Zipser 1985; Hertz et al. 1991; Rolls and Treves 1998; Rolls and Deco 2002). The weight vector points toward the center of the set of stimuli in the category. The different training stimuli that are placed into the same category (i.e. activate the same neuron) are typically

overlapping in that the pattern vectors are correlated with each other. Figure 3a illustrates this.

For the formation of invariant representations (by e.g. trace or CT learning), there are multiple occurrences of an object at different positions in the space. The object at each position represents a different transform (whether in position, size, view etc) of the object. The different transforms may be uncorrelated with each other, as would be the case for example with an object translated so far in the space that there would be no active afferents in common between the two transforms. Yet we need these two orthogonal patterns to be mapped to the same output. It may be a very elongated part of the input space that has to be mapped to the same output in invariance learning. These concepts are illustrated in Fig. 3b. In this paper, we show how continuous transformation learning as well as trace learning can contribute to solving translation invariance learning. Continuous transformation learning uses spatial continuity among the nearby positions of the object training exemplars to help produce the invariant object representations required. We emphasize that the network architecture and training rule can be identical for standard competitive learning and for continuous transformation learning. It is the training and style of learning that differs, with standard competitive learning pointing a synaptic weight vector toward the prototype of a set of training patterns, whereas continuous transform learning maps a continuous part of an elongated input space to the same output.

In conventional competitive learning, the overall weight vector points to the prototypical representation of the object. The only sense in which after normal competitive training (without translations etc) the network generalizes is with respect to the dot product similarity of any input vector compared to the central vector from the training set



**Fig. 3** **a** Conventional competitive learning. A cluster of overlapping input patterns is categorized as being similar, and this is implemented by a weight vector of an output neuron pointing toward the center of the cluster. Three clusters are shown, and each cluster might after training have a weight vector pointing toward it. **b** Invariant representation learning. The different transforms of an object may span an elongated region of the space, and the transforms at the far ends of the space may have no overlap (correlation), yet the network must learn to categorize them as similar. The different transforms of two different objects are represented

that the network learns. Continuous transformation learning works by providing a set of training vectors for each object that overlap and between them cover the whole space over which an invariant transform of the object must be learned. Indeed, it is important for continuous spatial transformation learning that the different exemplars of an object are sufficiently close that the similarity of adjacent training exemplars is sufficient to ensure that the same postsynaptic neuron learns to bridge the continuous space spanned by the whole set of training exemplars of a given object. This will enable the postsynaptic neuron to span a very elongated space of the different transforms of an object.

#### Simulations: stimuli

The stimuli used to train the networks were  $64 \times 64$  pixel images, with 256 gray levels, of frontal views of faces, examples of which have been shown previously (Wallis and Rolls 1997). In some of the experiments described here, just two face stimuli were used, as these are sufficient to demonstrate some of the major properties of CT learning.

#### Simulations: training and test procedure

To train the network, each stimulus is presented to the network in a sequence of different transforms, in this paper, positions on the retina. At each presentation of each transform of each stimulus, 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 number of training epochs for layers 1–4 was 50. The learning rates  $\alpha$  in Eqs. 5 and 6 for layer 1 were  $3.67 \times 10^{-5}$ , and for layers 2–4 were  $1.0 \times 10^{-4}$ .

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 translation 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 position. 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 in Rolls et al. (1997a) and Rolls and Milward (2000). 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 CT learning to train the network to recognize different face stimuli in different positions. The maximum single cell information measure is

$$\text{Maximum single cell information} = \log_2(\text{number of stimuli}). \quad (8)$$

For two face stimuli, this gives a maximum single (and multiple) cell information measure of 1 bit, and for four face stimuli 2 bits. The single cell information is maximal if, for example, a cell responds to one object (i.e. stimulus) in all locations tested and not to the other object or objects in any location. The multiple cell information is maximal if all objects have invariant neurons that respond to one but not the other objects.

## Results

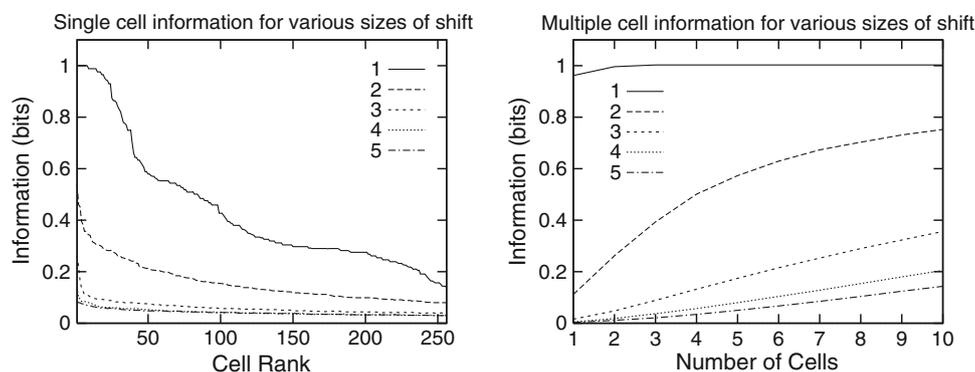
### Experiment 1: the effect of stimulus spacing

The aim of this experiment was to demonstrate translation invariance learning using a Hebb learning rule with no temporal trace. It was predicted that the CT effect would break down if the distance between nearest transforms was increased, as this would break the spatial continuity between the transforms of stimuli. This experiment investigated whether this occurred and how great the shift needed to be for the type of stimuli used before the CT learning started to fail.

Networks were created and trained with two test faces, on an  $11 \times 11$  square grid of training locations. Different distances between each training location were used in different simulations. The distances between locations in different simulations were 1, 2, 3, 4, and 5 pixels. Networks were trained with the Hebb rule (Eq. 5).

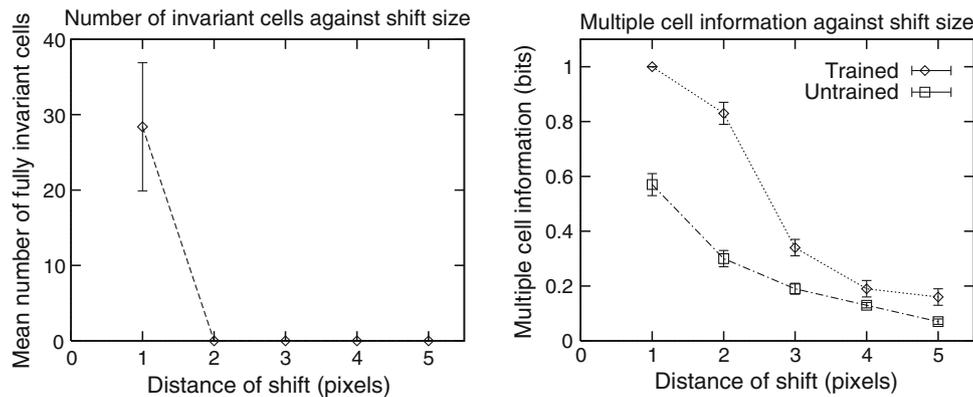
Figure 4 shows the performance for different spacings measured with the single and multiple cell information measures. The single cell measure shows for the 256 most invariant cells in layer 4 (which contains 1,024 cells) how much information each cell provided about which of the two stimuli was shown, independently of the location of the stimuli. Thus, a cell with one bit of information discriminated perfectly between the two stimuli for every one of the  $11 \times 11 = 121$  training locations, responding for example to one object in all 121 locations and not responding to the other object in any location. The multiple cell information shows that different neurons are tuned to each of the stimuli used. Both parts of Fig. 4 show that the performance decreases as the distance between adjacent training locations increases.

We now consider the average results of five simulation runs in each spacing condition, together with control results. (Each simulation run had different random seeds for the connectivity and connection weights.) For each simulation, the number of fully invariant cells in the output layer (i.e. those with 1 bit of stimulus-specific information) was averaged across all networks within a condition. These results are plotted in the left-hand panel of Fig. 5. Given that no invariant cells were found in any of the networks in the untrained condition, it is clear that the network showed a strong improvement after training when stimulus locations were only one pixel apart. However, when the separation between training locations was increased to 2 pixels spacing or more (i.e. 3, 4 and 5), no fully invariant cells were found.



**Fig. 4** *Left* Single cell information measures showing the performance of layer 4 neurons after training with the Hebb rule, with separate curves for networks trained with different distances (in pixels) between the training locations. The performance of the 256

most selective cells in layer 4 is shown, and perfect performance is one bit. *Right* The multiple cell information measures shows the performance of ensembles of neurons with the different distances between training locations



**Fig. 5** *Left* Plots of the mean number of fully invariant cells (i.e. cells with 1 bit of stimulus-specific information) averaged across all five networks after training over varying amounts of separation between training locations. *Right* Plots of the mean cumulative multiple cell information averaged across all five networks both before ('Untrained') and after ('Trained') training with varying amounts of

separation between training locations. Multiple cell information was measured across the ten best cells in the output layer of each network, which were selected by rank ordering the cells based on their stimulus-specific information for each stimulus in turn, and choosing the five best for that stimulus

The right-hand panel of Fig. 5 shows the mean cumulative multiple cell information contained in the responses for the ten best cells (i.e. five cells for each stimulus that contain the most stimulus-specific information about that stimulus) averaged across the five networks for each separation. The plot shows results from both trained and untrained (control) networks and shows that the trained network performs well with training locations spaced 1 pixel apart, moderately with the training locations spaced 2 pixels apart and much less well if the training locations are 3 or more pixels apart. Consistent with this, in the control untrained condition with merely random mappings, although some groups of cells do provide some information about which stimulus was shown, the amount of information for in particular one and two pixel spacings for the test locations is very much less than with training.

These results demonstrate that continuous transformation learning, which uses only a Hebb rule, can produce translation invariant representations in a hierarchical model of visual processing. When training locations are separated by a single pixel, the networks are clearly able to produce a number of cells that respond to the stimuli invariantly of location. That this effect is due to CT learning can be demonstrated by the fact that as the training locations are moved further apart performance dramatically decreases. The reason for the decrease in performance with two or more pixel spacings for the training locations is that the same postsynaptic neuron is not likely to be active for adjacent training locations, as described in Fig. 2. The results also indicate that for a  $64 \times 64$  image of a face, the CT training should ideally encompass adjacent locations for the training and that the training locations should not be more than 2 pixels apart.

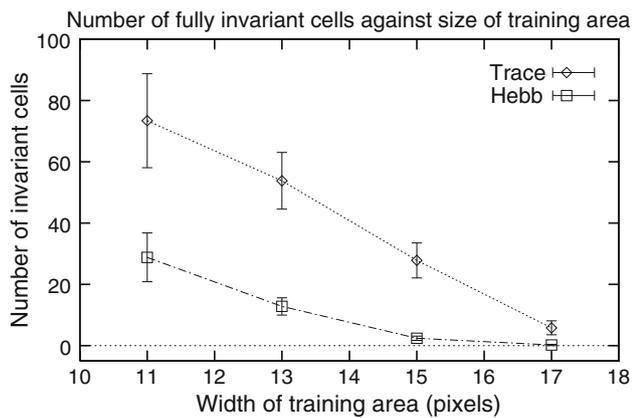
#### Experiment 2: how many locations can be trained?

The aim of the next simulations was to investigate how CT learning operates as the number of training locations over which translation invariant representations of images must be formed increases.

A series of networks was trained with stimuli presented on grids of training locations of various sizes. Each network was trained on the two face stimuli using  $11 \times 11$ ,  $13 \times 13$ ,  $15 \times 15$ , and  $17 \times 17$  grids of training locations. Given the findings of experiment 1, the training locations were a single pixel apart. Networks trained with the Hebb rule in the CT condition, and with the trace rule, were compared. As in experiment 1, five networks were created with different random seeds and run on each condition.

The mean number of cells with fully translation invariant representations (i.e. those with 1 bit of information with the single cell information measure) in the output layer of each network is shown in Fig. 6. Prior to training, no network had any cells with full translation invariance. Hence, all invariant cells in trained networks were due to learning. The results with the Hebb rule training condition shown in Fig. 6 in the lower curve show that with 121 training locations (the  $11 \times 11$  condition), training with CT only produces many fully translation invariant cells in layer 4 and that the number decreases gradually as the number of training locations increases. Even with 169 training locations (the  $13 \times 13$  condition), many cells had perfect translation invariance. With 225 ( $15 \times 15$ ) or more training locations, training with the Hebb rule produced very few cells with perfect translation invariance.

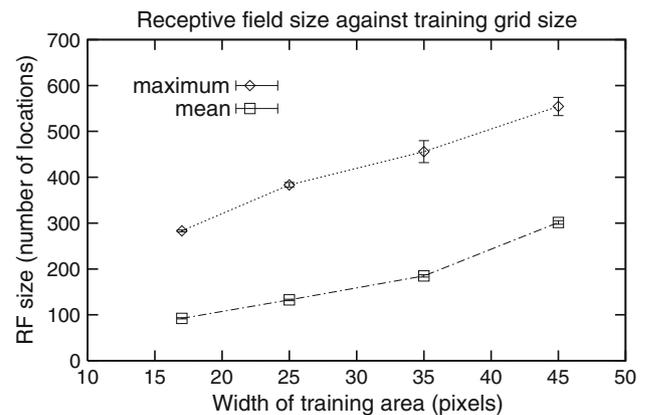
Figure 6 shows that training with the trace rule produced better performance than with the Hebb rule. Training with



**Fig. 6** Plots of the mean ( $\pm$ sem) of the number of fully invariant cells measured from the output layers of five networks trained with either the Hebb rule or the trace rule (Trace), with varying sized grids of training locations. The grids were square, and the width of each grid is shown. There were thus 225 training locations for the grid of width 15

the trace rule (provided that there is a small separation between training locations) allows the network to benefit from the CT effect, because the trace rule has a Hebbian component, and adjacent training locations are likely to activate the same postsynaptic neuron because of similarity in the input representations when the images are one pixel apart. But in addition, the trace rule shows that the additional effect of keeping the same postsynaptic neuron eligible for learning by virtue of closeness in time of exemplars of the same training object does help the network to perform better than with Hebb rule training, as shown in Fig. 6. The value of the trace parameter  $\eta$  was set in the above simulations to 0.8. This produces a drop in the trace value to  $1/e$  (37%) in 4–5 trials. We systematically explored whether longer traces would be useful in the simulations in which the 121 locations were presented (in 11-long sequences permuted from the 121 transforms) and showed that somewhat better performance than that illustrated in Fig. 6 could be produced by increasing the value of  $\eta$  up to 0.95. For example, with  $\eta = 0.95$ , which corresponds to trace decay to 37% in approximately 20 trials, the mean number of fully invariant cells (with a width of the training area of 11) was 83, compared to 73 with  $\eta = 0.8$ .

The aforementioned results were for cells that discriminated perfectly between the training images over all training locations. We now investigate how much larger the training areas can be (and thus the receptive field sizes) if the criterion of perfect discrimination (i.e. for every trained location) is relaxed. We used as a measure the number of locations at which the response to the preferred stimulus for a neuron (as indicated by the maximal response at any location) was more than a criterion amount above the



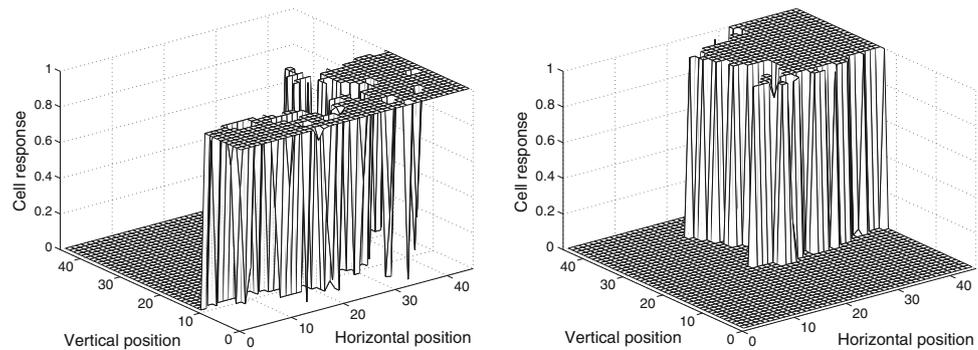
**Fig. 7** The RF size of neurons as a function of the width of the square training grid. The upper curve shows the mean ( $\pm$ sem) of the maximum receptive field size of a layer 4 selective neuron across 5 simulation runs. The lower curve is the mean ( $\pm$ sem) receptive field size of all selective layer 4 neurons across 5 simulation runs. The measure of selectivity is given in the text

largest response to the other stimulus at any location. This was designated as the size of the cell's receptive field (RF). The results did not depend closely on the exact value of the criterion amount (with values of 0.1 and 0.01 producing similar results), and the criterion amount was 0.01 for the results shown. (The scale is that the firing rate of the neurons is in the range 0.0–1.0.) The sizes of the receptive fields of neurons after training with up to 2,025 ( $45 \times 45$ ) locations are shown in Fig. 7. The largest receptive field found was 605 pixels in size, and the mean value with training with  $45 \times 45$  locations was approximately 300 locations. (605 pixels for a square receptive field would correspond to a grid of approximately  $24 \times 24$  pixels, which is a relatively large translation for a  $64 \times 64$  image.) (In the untrained control condition, with grid sizes of  $25 \times 25$  or greater, there were almost no layer 4 cells that showed selectivity by the above criteria.) Examples of the receptive fields found for layer 4 neurons are shown in Fig. 8 and include many adjacent locations because of the nature of continuous transformation learning illustrated in Fig. 2.

### Experiment 3: effects of training order

In principle, continuous transformation learning need not have the nearest exemplars of any image presented in any particular order, as it is the similarity in the inputs produced by the spatially closest images that drives the learning (see Fig. 2) rather than the temporal order of presentation. We tested this by performing simulations in which the complete set of training locations used in any epoch were presented in a permuted sequence for one image and then for the other. This is termed the 'permuted'

**Fig. 8** Mesh plots of the receptive fields of two output cells from networks trained on a  $45 \times 45$  grid of locations. The z-axis gives the cell response at a given training location



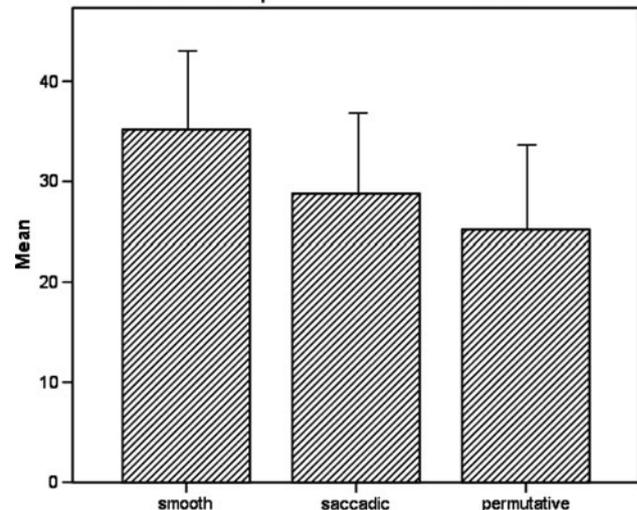
condition and is compared with the smooth movement along successive rows that is the default ('smooth') training condition. We also compared these training conditions with a 'saccadic' condition in which 11 locations were presented smoothly, followed by a jump to a new location for a further set of 11 smooth transitions, etc. (In the saccadic algorithm, care was taken to ensure that every location was visited once in every training epoch.) These permuted and saccadic training conditions were investigated not only for their theoretical interest, but also because they provide a way of systematically investigating the types of training condition that may occur in real-world environments, in which saccadic eye movements are usually made.

To test this, three sets of simulations were run in which five networks were trained with the two face stimuli over a  $11 \times 11$  grid of training locations with one pixel spacing for the training locations. The mean number of fully invariant cells (i.e. representing 1 bit of stimulus-specific information) in the output layer is shown in Fig. 9. The figure shows that though more invariant cells are produced on average in the 'smooth' condition, any difference between the three conditions is well within the standard error of both sets of data, and the networks trained with 'saccadically' and 'permuted' training conditions still continue to produce a large number of fully invariant cells. A one-way repeated-measures ANOVA (where the random seed used to initialize network weights and connections is the random factor) revealed that there was no significant difference between the three conditions ( $F(2,12) = 0.395$ ;  $p = 0.682$ ). Thus, these simulations show that the temporal order of presentation is not a crucial factor in whether the networks can be successfully trained to form translation invariant representations by CT learning. The implications of this are considered in the Discussion.

#### Experiment 4: how many stimuli can be trained?

The number of locations over which invariance can be trained is only one measure of the success of the model. While it is important that the network should respond

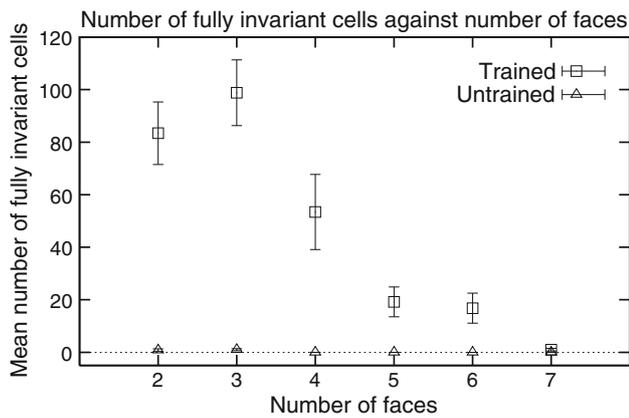
**Mean number of fully invariant cells produced by different presentations**



**Fig. 9** The mean number of fully invariant cells found in the output layer of networks trained with one of three methods of stimulus presentation (bars display the standard error of the mean). In the 'smooth' condition, stimuli always moved to a neighboring location at every timestep. In the 'saccadic' condition, the stimuli made periodic jumps in location. In the 'permuted' condition, the stimuli permuted through the locations in random order

invariantly over as many locations as possible, it is also of importance to investigate how many different stimuli it is capable of discriminating with translation invariance. The networks presented thus far have been trained to discriminate between only two faces. This is something of a simplification relative to previous studies of translation invariance in VisNet, where a minimum of seven faces have been used (see, for instance, Wallis and Rolls 1997; Rolls and Milward 2000). However, in those investigations of translation invariance learning with the trace rule, the distance between adjacent training locations was typically large (e.g. 32 pixels) (a condition in which continuous transformation learning will not operate), and the number of trained locations was for example nine.

In experiment 4, we investigated to what extent translation invariant representations of more than two stimuli



**Fig. 10** Capacity. The mean number of cells with the maximum single cell information when trained with different numbers of faces on a  $5 \times 5$  translation grid. The mean ( $\pm$ sem) across 5 simulation runs is shown. The performance for the untrained network is shown as a control

could be learned under conditions in which CT learning can operate as shown earlier, using a distance between the training locations of one pixel. We chose to use a  $5 \times 5$  grid of training locations, and in addition to the two faces used already, more faces taken from the set illustrated by Wallis and Rolls (1997). We compared the results of Hebb rule learning with the untrained control condition.

The results in Fig. 10 show that many cells have the maximum value of the single cell information about the face shown when the network is trained on a problem with 25 locations (in a  $5 \times 5$  grid) for up to four faces. To obtain the maximum value, a neuron had to respond to one of the faces better than to any other at every one of the 25 training locations. We checked that for the experiments with 2 or 3 faces, there were neurons that were tuned in this way to each of the faces. Beyond this, some cells still had the maximum value of the single cell information when trained on 4, 5, or 6 faces, but performance had started to decrease, and was low for 7 faces (see Fig. 10).

These findings show that the networks were able to learn to discriminate perfectly up to 3 faces for every one of 25 training locations for each face and that performance started to decrease with larger numbers of stimuli.

## Discussion

The results described here show that continuous transformation learning can provide a basis for forming translation invariant representations in a hierarchical network modeling aspects of the ventral visual stream. The evidence that continuous transformation learning was the process responsible for learning is that invariant representations were produced when the training rule was purely

associative (Hebbian) (Eq. 5), and that the learning was strongly impaired if similarity between the nearest exemplars was reduced by increasing the spacing between the training locations of the  $64 \times 64$  pixel face images by more than two pixels (Experiment 1). The exact number of pixels at which the CT effect breaks down will depend on the statistics of the images used, with images in which there are only low spatial frequencies predicted to allow larger training separations for the nearest locations. CT learning of view invariant representations has been established (Stringer et al. 2006), and this is the first demonstration of its use for learning translation invariant representations of objects. It is important to investigate the application of continuous transformation learning to translation invariance learning, for this is inherently more difficult than the view invariance in one axis that has been studied previously (Stringer et al. 2006; Perry et al. 2006). Translation invariance is more difficult in that the training needs to encompass shifting of the features in two dimensions,  $X$  and  $Y$ . That is, any one feature that is part of an object can occur in any one of the grid coordinates in two dimensions in which the 2D translation invariance is being learned. Two-dimensional translation invariance is what was learned in all of the Experiments described in the paper. Translation invariance is crucial to useful shape recognition in biological systems, and it is therefore of fundamental importance to verify that CT learning can solve this class of problem.

The results show that the number of training locations over which translation invariant representations can be produced using CT learning is quite large, with evidence for translation invariant representations for up to 605 pixels in size in experiment 2. (605 pixels for a square receptive field would correspond to a grid of approximately  $24 \times 24$  pixels).

The capacity with CT training for large numbers of training locations is of interest in relation to previous investigations with the trace rule, in which performance was good with 7 faces for 9 locations, but decreased markedly as the number of training locations was increased to 25 locations (Rolls and Milward 2000). However, in the previous investigations with the trace rule, the spacing between nearest training locations was large, e.g. 32 pixels. In the results described here, we have found that CT learning can learn with large numbers of training locations (Experiment 2) with two faces. In experiment 4, we showed that with 25 training locations, the number of stimuli that can be trained to have perfect translation invariant responses is 3, with performance starting to decline with more faces. On the other hand, the trace rule can cope with more stimuli, provided that the number of training locations is not very large (e.g. 9). What factors might underlie these performance differences?

CT translation invariance training relies on close training locations, as otherwise the similarity in the input representations is too small to support similar firing of the same postsynaptic neurons by the same stimulus at adjacent locations (Experiment 1). We note that low spatial frequencies will tend to reflect spatial continuity as an object transforms and may therefore be especially useful in supporting the operation of CT learning. In this context, it is of interest that the low-pass spatial frequency filtering that happens early during development could in fact be useful in helping representations built using the CT effect to be set up. Indeed, in the initial development of the training protocols used with CT learning, we explicitly trained with only the two low spatial frequencies present and found that this could facilitate CT learning (although all the investigations reported here and elsewhere (Stringer et al. 2006; Perry et al. 2006) used all four spatial frequency filter banks, as this can potentially help discrimination between objects).

One factor that may limit the performance of CT translation invariance learning is that if neurons in the network learn to represent a feature in an image such as an eye, then it is possible that a similar feature in another face will become spatially aligned with the feature from the first face when that image is moved continuously across the retina during training. Those neurons would then learn to respond to the same feature in different faces, and this would limit the discrimination capacity. (This is not a problem with trace learning when the step size is large, as then it is unlikely that two similar features from two different faces will overlap and lead to a CT association.) One possible solution to this issue in hierarchical object recognition networks is to learn rather specific feature combination representations in an early layer, so that the feature combination neurons will be different for different objects. CT learning (and also trace learning) could then learn translation invariant representations of the feature combination representations that are now different for the different faces / objects. It will be interesting to explore to what extent training in this regime will allow invariant representations to be formed using CT training for larger numbers of objects.

From the investigations described here and elsewhere, it appears that CT learning is useful under conditions where training images are available with small continuous differences between images, and under these conditions CT learning supports the learning of many transforms of objects. On the other hand, trace learning can operate well with larger numbers of training objects (up to e.g. 7, Rolls and Milward (2000)), though with fewer locations (with up to 25 tested). We suggest that a combination of these two learning processes could 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 surfaces 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 catastrophically different views. Further, temporal contiguity, a property of the transforms of real objects in the world, may 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 may in this case help to break apart the representations of different objects.

It would be of interest to explore how continuous spatial transformation and temporal trace training may complement each other further. It would, for example, be useful to investigate how the capacity of the network scales up with size when trained with each approach and with a combination of both. One point is that the stimulus-specific information or ‘surprise’ conveyed by neurons in VisNet (e.g. 1 bit as shown in Fig. 4) is in the same range as that in the macaque inferior temporal visual cortex, in which the mean was 1.8 bits, though the exact value depends greatly on the number of stimuli in the set, and on how sparse vs distributed the representation is (Rolls et al. 1997b). The representational capacity of individual neurons in VisNet and the macaque inferior temporal visual cortex is thus not inherently different. What does differ is the size of the system. In the macaque visual system, if the inferior temporal cortex had an area of 150 mm<sup>2</sup> in each hemisphere, a thickness of 2 mm, and a cell density of 20,000 neurons per mm<sup>3</sup> (Rolls and Deco 2002), this would imply that IT cortex contains in the order of  $1.2 \times 10^7$  neurons. The number of neurons in the final layer of VisNet is 1,024, so the real visual system is likely to be scaled up by at least an order of  $10^4$  times relative to VisNet. Depending on how VisNet does scale up, the suggestion is that it could perform invariant object recognition for large numbers of objects. Indeed, there appears to be no fundamental reason why the VisNet architecture will not scale up, in that when a related hierarchical architecture for invariant recognition is scaled up, it has been tested successfully with more than 1,000 images and is described as having performance as good as the best computer vision systems (Serre et al. 2007). To investigate this further, Rolls (in preparation) produced a new version of VisNet with 16 times as many neurons in each layer, so that the number of neurons in each layer was now  $128 \times 128$ , and the image presented was  $256 \times 256$  on a  $512 \times 512$  gray background. Consistent with this, the radius of the excitatory connection

feedforward topology shown in Table 1 was increased by 4. It has been demonstrated that the scaled-up version of VisNet can compute perfect translation invariant representations over at least 1,089 locations for 5 objects (100% correct, 2.32 bits of information for the single and multiple cell information analyses), which compares favorably with the 300 locations with 2 stimuli and 25 locations for 3 faces reported in this paper for VisNet with the standard size shown in Table 1. We conclude that the demonstration of principle, that translation invariant representations can be formed by CT learning, described in this paper is not limited to the values found with standard VisNet but can scale up. Continuous spatial transformation learning is thus an interesting learning principle for translation invariant representations.

Another point of comparison that could influence how the system scales up is the sparseness of the representation. In the study described here, we set the sparseness as shown in Table 4. In further investigations, we showed that the performance of VisNet when trained using CT learning is robust with respect to the exact values of the sparseness used. For example, in additional simulations, we showed that the performance of the system was little affected by varying the sigmoid threshold parameter used in layer 1 from the standard value of 0.992 shown in Table 4 in the range 0.99 to 0.1. When the value was 0.1, 90% of the neurons in layer 1 had high activity to any one stimulus, yet good performance was found because a few cells were still well tuned to respond to only some of the stimuli, providing reasonable discrimination performance. Thus, CT learning does not appear to impose strong constraints on the sparseness values used in order to obtain good training.

We emphasize that continuous transformation learning is a different principle to methods that 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 and Rolls 1997; Rolls and Milward 2000; Rolls and Stringer 2001; Stone 1996; Bartlett and Sejnowski 1998; Becker 1999; Einhäuser et al. 2002; Wiskott and Sejnowski 2002). As discussed earlier, trace learning may usefully complement CT learning; and correspondingly, CT learning may be present when training with a trace rule if the training stimuli are sufficiently similar. It is shown for example in Fig. 6 that training with the trace rule produces better performance than with the Hebb rule. Part of the reason for this may be that, due to the diluted connectivity of the network, there may be some breaks in the spatial continuity of the input to cells at any stage of the network, and in these cases, by keeping the postsynaptic neurons eligible for learning on the short time scale, the trace rule may enable such discontinuities to be bridged.

Experiment 3 showed that translation invariant representations can be learned by CT learning when the order of stimuli is permuted, that is, the different views of an object occur in random temporal order during training. This emphasizes that it is spatial similarity that drives CT learning and that no temporal contiguity is necessary. In fact, even if visual fixation shifts rapidly and randomly to produce large translations of an object or face by saccades, CT learning can still develop invariant representations of the individual objects.

One prediction of continuous transformation learning is that invariant transform learning could occur under conditions when there is only spatial similarity between the training images, but temporal contiguity of images of the same object is broken by, for example, interleaving views of different objects during training. This has been tested psychophysically in humans, and it has been found that some view invariant learning can occur under these interleaved training conditions (Perry et al. 2006). We now make a corresponding prediction for the learning of translation invariant representations of objects. However, in the study by Perry et al. (2006), human learning was better if adjacent views of an object occurred in temporal succession, and this as well as other evidence (Wallis and Bulthoff 2001; Wallis 1998, 2002; Stone 1998) suggests that temporal continuity is a useful factor in helping humans to learn view invariant representations of objects. Indeed, we suggest that in the natural world temporal continuity is usually present with spatial continuity too. Thus, while the results in this paper establish that spatial continuity (as in CT learning) is sufficient to support translation invariance learning, we regard it as a mechanism that complements temporal continuity (as in trace) learning, with the strengths of each described elsewhere in this Discussion. Further, spatial continuity is usually present when there is temporal continuity, and that statistical fact about the natural world, together with the findings described in this paper, indicate that continuous spatial transformation learning is likely to play a role in the learning of invariant representations. We have been able to demonstrate that CT learning can usefully add to trace learning in the following simulations with the scaled-up version of VisNet (Rolls, in preparation). When training with the trace rule with 5 objects at 169 training locations, we found better invariance learning on a  $33 \times 33$  pixel grid (where the spacing is on average 2.5 pixels between training locations) than on a  $65 \times 65$  pixel grid (where the spacing is 5 pixels between training locations) ( $p < 0.004$  Mann–Whitney  $U$ -test,  $U = 0$ , with an average of 2.14 vs. 1.49 bits of information respectively about which of each of the stimuli was present). This shows that with 5 pixels between each training location, CT can make a smaller contribution to the training than when there are on average

2.5 pixels between each training location. This experiment thus shows that CT learning can usefully contribute to temporal trace learning when there is sufficient spatial continuity to allow the CT effect to operate.

An important aspect of the architecture of VisNet is that it can generalize to some transforms of an object that have not been trained. Indeed, part of Rolls' hypothesis (1992) is that training early layers (e.g. 1–3) with a wide range of visual stimuli will set up feature analysers in these early layers which are appropriate later on with no further training of early layers for new objects. For example, presentation of a new object might result in large numbers of low-order feature combination neurons in early layers of VisNet being active, but the particular set of feature combination neurons active would be different for the new object. The later layers of the network (in VisNet layer 4) would then learn this new set of layer 3 neurons active as the new object. However, if the new object was then shown in a new location, the same set of layer 3 neurons would be active because they respond with spatial invariance to feature combinations, and given that the layer 3–4 connections had already been set up by the new object, the correct layer 4 neurons would be activated by the new object in its new untrained location and without any further training. This principle was shown to apply for translation invariance learning with the trace rule (Elliffe et al. 2002) and with view invariance learning with continuous transform learning (Stringer et al. 2006). However, there are limitations to how much translation invariance can occur in the networks, and there is evidence that behaviorally translation invariance may be limited (Nazir and O'Regan 1990; Kravitz et al. 2008; McKyton et al. 2009), and that neurons in the primate inferior temporal visual cortex do not automatically generalize their responses to novel locations when learning new shapes (Cox and DiCarlo 2008). We were able to confirm that some generalization over location does occur with CT learning as follows. In the scaled-up version of VisNet (Rolls, in preparation), we trained 5 objects in 9 locations in translations in a coordinate range of  $13 \times 13$  pixels across the input. The training locations were at the center, the four cardinal points, and the four corners, of this coordinate range, i.e. in a  $3 \times 3$  arrangement, with an average translation of 4.3 pixels between each training location. We then tested in all 169 coordinates in that range (in 160 of which the image had not been placed before). We found that many cells in layer 4 discriminated the 5 stimuli perfectly at 100% correct across all 169 locations, by responding to one but to none of the other stimuli across all 169 locations with a single cell information value of 2.3 bits. In an untrained control, no cells discriminated perfectly between the stimuli, with most cells showing poor discrimination between the stimuli. Thus, translation

invariance learning using CT does show some generalization across position (across for example 4 pixels in the scaled-up version of Visnet), for reasons of the type described elsewhere.

In conclusion, the results described in this paper are the first to demonstrate the use of continuous spatial transformation learning to build translation invariant representations of objects in a hierarchical object recognition network. We further propose that continuous transformation learning could contribute to learning in any sensory system in the brain with continuity (e.g. spatial continuity) in the sensory space, for instance the somatosensory system, or the hippocampus (Rolls et al. 2006). Indeed, CT learning is very likely to operate in such systems given certain parameters, including a synaptic learning rate that is sufficiently high and continuous spatial variation of the sensory input. The continuous spatial transformation learning works because with the natural statistics of images in the world, and the spatial filtering performed in the visual system, images of a given object that are shifted by a small distance will tend to activate the same postsynaptic neuron, and it is this that enables continuous transformation learning to operate. We emphasize that this continuous spatial transformation learning can occur with no temporal continuity at all, as shown by invariance learning that can be specific for a particular object even when the different transforms of different objects are interleaved in the order in which they are presented.

**Acknowledgments** This research was supported by the Wellcome Trust and by the MRC Interdisciplinary Research Centre for Cognitive Neuroscience. GP was supported by the Medical Research Council.

## References

- 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 Appl* 11:347–374
- Biederman I (1987) Recognition-by-components: a theory of human image understanding. *Psychol Rev* 94(2):115–147
- Booth MCA, Rolls ET (1998) View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb Cortex* 8:510–523
- Cox DD, DiCarlo JJ (2008) Does learned shape selectivity in inferior temporal cortex automatically generalize across retinal position? *J Neurosci* 28:10045–10055
- Deco G, Rolls ET (2004) A neurodynamical cortical model of visual attention and invariant object recognition. *Vis Res* 44:621–644
- Deco G, Rolls ET (2005) Attention, short term memory, and action selection: a unifying theory. *Prog Neurobiol* 76:236–256
- Desimone R (1991) Face-selective cells in the temporal cortex of monkeys. *J Cogn Neurosci* 3:1–8
- Einhäuser W, Kayser C, König P, Körding KP (2002) Learning the invariance properties of complex cells from their responses to natural stimuli. *Eur J Neurosci* 15:475–486

- 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 Appl* 3:194–200
- 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
- Giese MA, Poggio T (2003) Neural mechanisms for the recognition of biological movements. *Nat Rev Neurosci* 4:179–192
- 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
- Hawken MJ, Parker AJ (1987) Spatial properties of the monkey striate cortex. *Proc R Soc Lond B* 231:251–288
- Hertz J, Krogh A, Palmer RG (1991) Introduction to the Theory of Neural Computation. Addison Wesley, Wokingham
- 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
- Kobatake 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 Press, Cambridge, Massachusetts
- Kravitz DJ, Vinson LD, Baker CI (2008) How position dependent is visual object recognition? *Trends Cogn Sci* 12:114–122
- McKyton A, Pertzov Y, Zohary E (2009) Pattern matching is assessed in retinotopic coordinates. *J Vis* 9(13):19 1–10
- Nazir TA, O'Regan JK (1990) Some results on translation invariance in the human visual system. *Spat Vis* 5:81–100
- Op de Beeck H, Vogels R (2000) Spatial sensitivity of macaque inferior temporal neurons. *J Comp Neurol* 426:505–518
- Perry G, Rolls ET, Stringer SM (2006) Spatial vs temporal continuity in view invariant visual object recognition learning. *Vis Res* 46:3994–4006
- Pinto N, Cox DD, DiCarlo JJ (2008) Why is real-world visual object recognition hard? *PLoS Comput Biol* 4:e27
- Riesenhuber M, Poggio T (1999a) Are cortical models really bound by the “binding problem”? *Neuron* 24:87–93
- Riesenhuber M, Poggio T (1999b) Hierarchical models of object recognition in cortex. *Nat Neurosci* 2:1019–1025
- Riesenhuber M, Poggio T (2000) Models of object recognition. *Nat Neurosci Suppl* 3:1199–1204
- Rolls ET (1992) Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philos Trans R 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 (2007) The representation of information about faces in the temporal and frontal lobes of primates including humans. *Neuropsychologia* 45:124–143
- Rolls ET (2008a) Face representations in different brain areas, and critical band masking. *J Neuropsychol* 2:325–360
- Rolls ET (2008b) Memory, attention, and decision-making. A unifying computational neuroscience approach. Oxford University Press, Oxford
- Rolls ET (2009) The neurophysiology and computational mechanisms of object representation. In: Dickinson S, Tarr M, Leonardis A, Schiele B (eds) Object categorization: computer and human vision perspectives, Chap. 14. Cambridge University Press, Cambridge, pp. 257–287
- Rolls ET (2010) Face neurons. In: Calder AJ, Rhodes G, Johnson MH, Haxby JV (eds) The handbook of face perception. Oxford University Press, Oxford
- Rolls ET, Aggelopoulos NC, Zheng F (2003) The receptive fields of inferior temporal cortex neurons in natural scenes. *J Neurosci* 23:339–348
- 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, 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 Appl* 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, Stringer SM (2006) Invariant visual object recognition: a model, with lighting invariance. *J Physiol Paris* 100:43–62
- Rolls ET, Stringer SM, Elliot T (2006) Entorhinal cortex grid cells can map to hippocampal place cells by competitive learning. *Netw Comput Neural Syst* 17:447–465
- Rolls ET, Treves A (1998) Neural networks and brain function. Oxford University Press, Oxford
- 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
- Rolls ET, Tromans JM, Stringer SM (2008) Spatial scene representations formed by self-organizing learning in a hippocampal extension of the ventral visual system. *Eur J Neurosci* 28:2116–2127
- Rumelhart DE, Zipser D (1985) Feature discovery by competitive learning. *Cogn Sci* 9:75–112
- Serre T, Oliva A, Poggio T (2007) A feedforward architecture accounts for rapid categorization. *Proc Nat Acad Sci* 104:6424–6429
- Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T (2007) Robust object recognition with cortex-like mechanisms. *IEEE Trans Pattern Anal Mach Intell* 29:411–426
- Stone JV (1996) Learning perceptually salient visual parameters using spatiotemporal smoothness constraints. *Neural Comput Appl* 8:1463–1492
- Stone JV (1998) Object recognition using spatiotemporal signatures. *Vis Res* 38:947–951
- Stringer SM, Perry G, Rolls ET, Proske JH (2006) Learning invariant object recognition in the visual system with continuous transformations. *Biol Cybern* 94:128–142
- 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 Press, 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 (1998) Temporal order in human object recognition. *J Biol Syst* 6:299–313
- Wallis G (2002) The role of object motion in forging long-term representations of objects. *Vis Cogn* 9:233–247
- Wallis G, Baddeley R (1997) Optimal unsupervised learning in invariant object recognition. *Neural Comput Appl* 9:883–894
- Wallis G, Bulthoff HH (2001) Effects of temporal association on recognition memory. *Proc Nat Acad Sci* 98:4800–4804
- Wallis G, Rolls ET (1997) Invariant face and object recognition in the visual system. *Prog Neurobiol* 51:167–194
- Wiskott L, Sejnowski TJ (2002) Slow feature analysis: unsupervised learning of invariances. *Neural Comput Appl* 14:715–770