Vision Research 119 (2016) 16–28

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

A computational exploration of complementary learning mechanisms in the primate ventral visual pathway

Courtney J. Spoerer *, Akihiro Eguchi *, Simon M. Stringer

Oxford Centre for Theoretical Neuroscience and Artificial Intelligence, Department of Experimental Psychology, University of Oxford, United Kingdom

ARTICLE INFO

Article history:
Received 11 March 2015
Received in revised form 21 September 2015
Accepted 8 December 2015

Keywords:
Visual object recognition
Continuous transformation
Trace learning
Inferior temporal cortex

ABSTRACT

In order to develop transformation invariant representations of objects, the visual system must make use of constraints placed upon object transformation by the environment. For example, objects transform continuously from one point to another in both space and time. These two constraints have been exploited separately in order to develop translation and view invariance in a hierarchical multilayer model of the primate ventral visual pathway in the form of continuous transformation learning and temporal trace learning. We show for the first time that these two learning rules can work cooperatively in the model. Using these two learning rules together can support the development of invariance in cells and help maintain object selectivity when stimuli are presented over a large number of locations or when trained separately over a large number of viewing angles.

1. Introduction

In vision, it is important to correctly identify an object in the environment as being the same despite changes in the retinal image. Over successive stages in the visual system, neurons develop response properties that are invariant to the size, position, and view of an object (Rolls, 1992; Rolls, 2000; Rolls & Deco, 2002; Desimone, 1991; Tanaka, Saito, Fukada, & Moriya, 1991). Cells in inferior temporal cortex (IT) that show invariance to the translation (Op de Beeck & Vogels, 2000; Kobotake & Tanaka, 1994; Ito, Tamura, Fujita, & Tanaka, 1995; Tovee, Rolls, & Azzopardi, 1994), size (Rolls & Baylis, 1986; Ito et al., 1995), contrast (Rolls & Baylis, 1986), lighting (Vogels & Biederman, 2002), spatial frequency (Rolls, Baylis, & Leonard, 1985; Rolls, Baylis, & Hasselmo, 1987), and view (Hasselemo, Rolls, Baylis, & Nalwa, 1989; Booth & Rolls, 1998) of objects have been reported.

Developing invariant recognition of objects involves associating together representations of the same object under different conditions. In the particular case of translation invariance, this would mean developing associations between the neural representations of an object in different spatial locations on the retina. In order to develop these associations, the visual system can exploit constraints placed upon object translation by the environment. For example, when an object translates from one point to another, it does so in a manner that is continuous in both space and time. These same constraints can be exploited for the development of view invariance, as different views of an object also appear in a spatially and temporally continuous manner.

One method for developing translation invariant representations utilizes the temporally continuous nature of object translation. Neurophysiological evidence suggests that the brain might use this type of information to develop translation invariant representations of objects (Li & DiCarlo, 2008). As breaking temporal continuity causes neurons to lose their selective responses to different objects. Different approaches have been developed in order to understand how the brain might exploit this temporally continuity, such as using inputs representing temporal context to guide learning (Becker, 1999), learning high probability sequences of visual input in order to infer the object being presented (George & Hawkins, 2005), and extracting slowly changing features in the visual inputs to analyze the transform invariant representations (Berkes & Wiskott, 2005; Wiskott & Sejnowski, 2002).

Temporal information can also be used to develop invariant representations of objects by incorporating a temporal trace into associative learning rules (Földiák, 1991; Rolls, 1992; Wallis & Rolls, 1997). This encourages neurons to respond to stimulus image transforms that occur close together in time. The advantage of this approach is that it can arise naturally out of biophysically realistic spiking neural networks when longer time constants for synaptic conductance are introduced (Evans & Stringer, 2012). Increasing this time constant keeps the neuron active for longer as it lengthens the time period over which current leaks into the

* Corresponding authors.
E-mail addresses: courtney.spoerer@mrc-cbu.cam.ac.uk (C.J. Spoerer), akhiro.eguchi@psy.ox.ac.uk (A. Eguchi).

http://dx.doi.org/10.1016/j.visres.2015.12.008
0042-6989/

postsynaptic neuron, thus allowing temporal trace learning to occur. Therefore, it is feasible that this type of learning could occur in the brain without requiring a specific architecture to operate.

A second method for developing translation and view invariance, known as continuous transformation (CT) learning, depends on the spatial continuity of object transformation (Stringer, Perry, Rolls, & Proske, 2006). As an object moves smoothly from one location to another, it will also appear in several intermediate positions. Each of these intermediate positions will be highly overlapping with the adjacent locations that the object appears in as it moves across the environment. Therefore, each of these adjacent locations would be likely to activate a common postsynaptic neuron that associates each of the positions together. This would result in the cell developing translation invariant response properties.

Each of the methods discussed so far consider how spatial and temporal constraints could each individually contribute to the development of invariant representations. However, in the real world, information provided by each of these constraints is available to the visual system simultaneously. Psychophysical evidence suggests that object-selective view-invariant recognition is improved when stimuli transform in a temporally and spatially continuous manner, compared to spatially continuous transformation alone (Perry, Rolls, & Stringer, 2006). It is important to understand how an observer might simultaneously utilize the benefits of spatial and temporal continuity in object transformation when developing invariant representations. This effect could be explained by the visual system using CT learning and temporal trace learning in tandem.

In this paper, we will explore how CT learning and temporal trace learning can operate together to help develop view and translation invariance using a hierarchical model of the ventral visual pathway, VisNet (Wallis & Rolls, 1997; Rolls & Milward, 2000), illustrated in Fig. 1. Both trace and CT learning have been tested extensively in the rate-coded VisNet model (Wallis & Rolls, 1997; Stringer et al., 2006), and so we shall use VisNet to study how these two learning mechanisms may be combined in the same rate-coded model.

2. Methods

2.1. The VisNet model

2.1.1. Hierarchical neural network architecture of the model

The architecture of the model used in this paper, VisNet (Wallis & Rolls, 1997), is developed according to the following principles:

(i) A series of hierarchical competitive networks with local graded inhibition and excitation.
(ii) Convergent connections to each neuron from a topologically corresponding region of the preceding layer.
(iii) Synaptic plasticity based on a biologically-plausible local learning rule, such as the Hebb rule or trace rule.

As mentioned above, the forward connections to individual cells in VisNet are derived from a topologically corresponding location in the preceding layer. The probability of each connection forming follows a Gaussian distribution. These distributions are defined by a radius containing approximately 67% of the connections from the preceding layer. The values employed in the current study are given in Table 1. The gradual increase in the receptive field of cells in successive layers reflects the known physiology of the primate ventral visual pathway (Freeman & Simoncelli, 2011; Pasupathy, 2006; Pettet & Gilbert, 1992).

2.1.2. Pre-processing of the visual input by Gabor filters

Before images are presented to layer 1 of VisNet, they are pre-processed by a set of Gabor filters that correspond to the known response profiles of V1 simple cells (Jones & Palmer, 1987; Cumming & Parker, 1999). Filtering the images produces a unique set of inputs that are then presented to layer 1 of the model. The input filters used are computed by the following equations:

$$g(x, y, \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{x'}{\lambda} + \psi\right) \tag{1}$$

with the following definitions:

$$\begin{aligned} x' &= x\cos\theta + y\sin\theta \\ y' &= -x\sin\theta + y\cos\theta \end{aligned} \tag{2}$$

where x and y specify the position of a light impulse in the visual field (Petkov & Kruizinga, 1997), σ controls the number of such periods inside the Gaussian window, θ defines the orientation of the feature, ψ defines the phase, and γ sets the aspect ratio that determines the shape of the receptive field. In each experiment, an array of Gabor filters is generated at each of 256×256 retinal locations with the parameters given in Table 2.

The outputs of the Gabor filters are passed to the neurons in layer 1 of VisNet according to the synaptic connectivity given in Table 1. Each layer 1 neuron received connections from 400 randomly chosen Gabor filters within a topologically corresponding region of the retina.

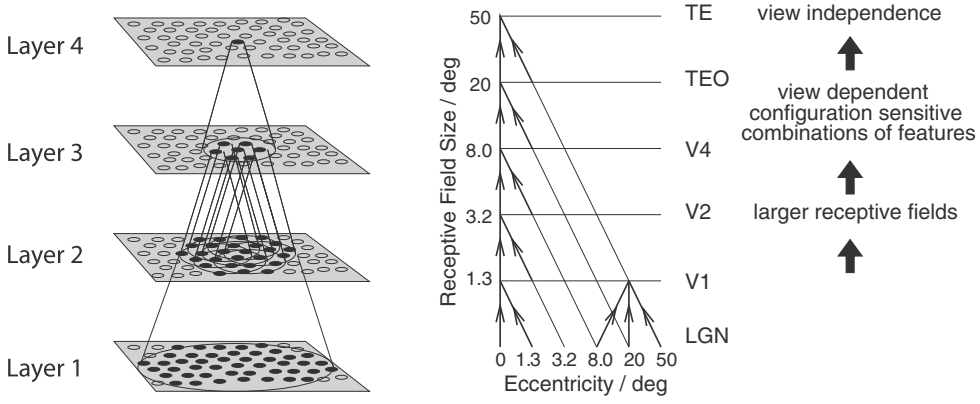

Fig. 1. (Left) A schematic representation of VisNet. The model consists of a hierarchy of competitive networks with feed-forward connections between them. Convergence in the model is designed so that cells in the final layer of the model have a receptive field that covers the whole of the input retina. (Right) Convergence in the visual system. V1, visual cortex area V1; TEO, posterior inferior temporal cortex; TE, inferior temporal cortex (IT).

Table 1
VisNet parameters.

Layer	Dimensions[a]	Number of connections[b]	Radius[c]
Layer 4	64 × 64	200	24
Layer 3	64 × 64	200	18
Layer 2	64 × 64	200	12
Layer 1	64 × 64	400	12
Retina	256 × 256 × 16		

Notes:
[a] Number of neurons per layer.
[b] Number of connections per neuron from the preceding layer.
[c] Radius from which 67% of the connections from cells in the preceding layer are received.

Table 2
Parameters for Gabor input filters.

Parameter (Symbol)	Value(s)
Phase shift (ψ)	0: white on black bar π: black on white bar
Wavelength (λ)	2, 16
Orientation (θ)	0, $\pi/4$, $\pi/2$, $3\pi/4$
Spatial bandwidth (b)	1.5 octaves
Aspect ratio (γ)	0.5

2.1.3. Calculation of cell activations within the network

For each of the cells in layers 1 to 4 of VisNet, the activation h_i of each neuron i was set equal to the linear sum of the inputs y_j from afferent neurons j in the preceding layer weighted by the synaptic weights w_{ij} as follows:

$$h_i = \sum_j w_{ij} y_j \tag{3}$$

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.

2.1.4. Interactions within layers

In these experiments, we ran simulations with a self-organizing map (SOM), (Von der Marlsburg, 1973; Kohonen, 1982) implemented within each layer. In the case of the SOM architecture, short-range excitation and long-range inhibition are combined to form a Mexican-hat spatial profile and is constructed as a difference of two Gaussians as follows:

$$I_{a,b} = -\delta_I \exp\left(-\frac{a^2+b^2}{\sigma_I^2}\right) + \delta_E \exp\left(-\frac{a^2+b^2}{\sigma_E^2}\right) \tag{4}$$

To implement the SOM, the activations h_i of neurons within a layer were convolved with a spatial filter, I_{ab}, where δ_I controlled the inhibitory contrast and δ_E controlled the excitatory contrast. The width of the inhibitory radius was controlled by σ_I and the width of the excitatory radius by σ_E. The parameters a and b indexed the distance away from the center of the filter. The lateral inhibition and excitation parameters used in the SOM architecture are given in Table 3.

Table 3
SOM parameters.

Layer	1	2	3	4
Excitatory radius (σ_E)	1.4	1.1	0.8	1.2
Excitatory contrast (δ_E)	5.35	33.15	117.57	120.12
Inhibitory radius (σ_I)	2.76	5.4	8.0	12.0
Inhibitory contrast (δ_I)	1.6	1.5	1.5	1.5

2.1.5. Contrast enhancement of neuronal firing rates

Next, the contrast between the activities of neurons within each layer was enhanced by passing the activations of the neurons through a sigmoid transfer function (Rolls & Treves, 1998) as follows:

$$y = f^{sigmoid}(r) = \frac{1}{1+\exp(-2\beta(r-\alpha))} \tag{5}$$

where r is the activation after applying the SOM filter, 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 within each layer of neurons to control the sparseness of the firing rates. For example, to set the sparseness to 4%, the threshold is set to the value of the 96th percentile point of the activations within the layer.

With the recent advances in computational capabilities, it is now easy to simulate visual pathway with more physiologically accurate spiking neural network models such as a conductance-based leaky integrate-and-fire neuron model (Evans & Stringer, 2013) or a Hodgkin–Huxley model (Eguchi, Neymotin, & Stringer, 2014). Therefore, some consider that the use of the sigmoid transfer function is no longer justifiable. However, the sigmoid transfer function is still the standard activation function used in rate-coded neural network modeling of brain function as they represent the fact that the firing rates of neurons are bounded and it introduces non-linearities to the network, making them an appropriate choice for the model (Ranzato, Huang, Boureau, & LeCun, 2007; Erhan et al., 2010; Ngiam et al., 2011). Furthermore, despite the complex mechanisms that cause a neuron to fire (Mainen et al., 1995; Destexhe & Pare, 1999), it has been claimed that a much simpler sigmoid activation function can provide a reasonable approximations at the level of population dynamics as the average of many different threshold functions becomes nonlinear (Marreiros, Daunizeau, Kiebel, & Friston, 2008). The parameters we used for the sigmoid activation function are those shown in Table 4, which have previously been selected after a number of optimization runs (Rolls, 2007).

2.1.6. Parameter setting

The parameter settings for these simulations were based upon values that optimized invariance learning in the network in previous experiments (Rolls, 2007; Tromans, Harris, & Stringer, 2011). This is with the exception of sparseness, which controls α in Eq. (3); the parameters were chosen based on the values that optimized performance on invariance learning in terms of the single cell information about each face in the current simulations. The justification for this exception is that sparseness levels are unlikely to be constant in the visual system and different sparseness levels are likely to be optimal for particular tasks (Rolls & Tovee, 1995). In the current simulations, sparseness levels were kept constant for both the translation invariance and view invariance simulations, and the high sparseness in the earlier layer reflects the physiological observations in some respect (Vinje & Gallant, 2000; Olshausen & Field, 2004).

2.1.7. Learning rules

For these simulations, two different learning rules were used to modify the strength of feed-forward synaptic connections between

Table 4
Parameters for the sigmoid activation function.

Layer	1	2	3	4
Percentile	98	90	90	90
Slope (β)	190	40	75	26

neurons within the network. The first rule used was the Hebb rule, where synaptic weights are updated by the following rule:

$$\Delta w_j = \alpha y x_j \tag{6}$$

where x_j is the firing rate of the jth presynaptic neuron, α is the learning rate (set in the interval between 0 and 1), y is the firing rate of the postsynaptic neuron, and w_j is the synaptic weight of the jth input to the postsynaptic neuron.

The second rule used was the trace rule. This rule is similar to the Hebb rule described above (Eq. (6)). However, it incorporates a trace of recent neuronal activity in the postsynaptic term, $\bar{y}^\tau$, at time step τ. The rule has the effect of encouraging the postsynaptic neuron to respond to input patterns that occur close together in time during training. The standard form of the trace rule is given by the following rule:

$$\Delta w_j = \alpha \bar{y}^\tau x_j^\tau \tag{7}$$

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

$$\bar{y}^\tau = (1-\eta) y^\tau + \eta \bar{y}^{\tau-1} \tag{8}$$

The parameter η is set in the interval between 0 and 1.

However, in this study a variant of the trace rule was used where the trace activity is taken from the immediately preceding time step, so the rule becomes

$$\Delta w_j = \alpha \bar{y}^{\tau-1} x_j^\tau \tag{9}$$

This variant was used as it has been shown to improve upon the performance of the standard trace rule in developing transform invariant representations (Rolls & Milward, 2000).

2.2. Training procedure

The synaptic weights in the network were initiated with random values. The simulations were repeated three times, each using a different random seed. This was done to ensure that the observed performance was not simply a consequence of the initial synaptic weights.

During training, an image of each face was presented to the network. The images were initially pre-processed by the Gabor input filters, and the output from these filters was used as input to the first layer of VisNet. Each cell in the first layer received a combination of inputs from 400 randomly chosen Gabor filters. The activation is then propagated through the network, using Eq. (3) to calculate individual cell activations and Eqs. (4) and (5) to determine firing rates. The weights are then updated according to either the Hebb rule (Eq. (6)) or the trace rule (Eq. (9)) depending on the condition.

For each face, images corresponding to successive transforms are presented sequentially to the network. For example, in the translation invariance simulations, the next image is the same as the previous image, but shifted by one pixel. The network then updates the synaptic weights after each individual image presentation. This process is repeated for each different face.

2.3. Stimuli

Faces were chosen as the stimuli to use in the simulations as they are complex stimuli that would make invariance learning more difficult. Increasing the difficulty of the task reduces the risk of the network reaching ceiling performance, which would make it difficult to see any difference that might exist between the two learning rules. This should enhance any beneficial effects of the learning rules, making these effects clearer to identify. The faces were generated using FaceGen 3D face modeling software (an example is shown in Fig. 2).

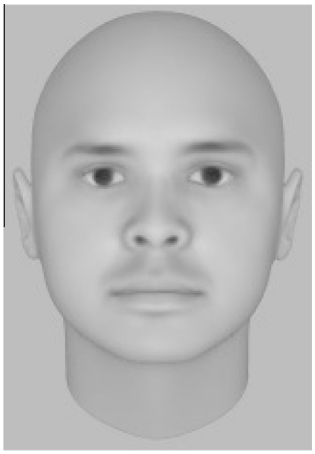

Fig. 2. An example of the stimuli used in the experiment.

The number of different positions or viewing angles were varied across three different experiments. For all experiments, each face was presented so that the stimulus transformed smoothly across space (Fig. 3), with no more than a 1 pixel shift between each presentation in the translation invariance experiments, and with no more than 1° change in viewing angle in the view invariance experiments, as is required for continuous transformation learning to operate (Stringer et al., 2006). The faces were also presented with temporal continuity, so that each successive transform of a face is presented in sequence before showing the transforms of the next face.

2.4. Information based measures of performance

To assess performance, a single cell measure of information was used for cells in the final layer (layer 4) of the model (Rolls & Milward, 2000). This measure describes the amount of information, $I(s,R)$, that a set of responses, R, gives about a stimulus, s, e.g. a particular face.

$$I(s,R) = \sum_{r \in R} P(r|s) \log_2 \frac{P(r|s)}{P(r)} \tag{10}$$

where r is an individual response from the set of responses, R, of a particular neuron. Information is highest when a cell responds invariantly to a stimulus across all different locations or views, but does not respond to other stimuli. We can then calculate the maximum information, I_{max}, that a cell can have with the following formula:

$$I_{max} = \log_2(N) \tag{11}$$

where N is the number of different stimuli s. We computed the maximum amount of information a cell conveyed about a particular face identity, using the same stimuli as in testing. This is as opposed to mutual information, $I(S,R)$, where S is the whole set of different stimuli s. Examples of the cell information calculations are given in Appendix A.

3. Results

In the experiment 1, we investigated the development of translation invariant representation of faces and compared the performance between the network when the synaptic connections were updated based on either the Hebb rule or the trace rule. In the experiment 2, we investigated the development of rotational view invariant representations of faces, and finally, in the experiment 3, we

Fig. 3. An example of a single face from different viewing angles. Each image is separated by 20°.

investigated the development of both the translation and rotational view invariant representations of faces.

3.1. Experiment 1: Translation invariance

3.1.1. Model performance with 4 faces in 100 locations

Experiment 1 compared the performance of the model when trained with the two different learning rules using a stimulus set containing a small number of faces in a large number of retinal locations (Fig. 4). In this scenario, the network was trained with 4 faces in 100 different retinal locations ($N = 4,\ I_{max} = 2$). Mann–Whitney U tests were used to test for differences in single-cell information measures between learning rules in all experiments.

In this experiment, single cell information measures were significantly lower in the untrained model ($Mdn = 0.30$) compared to the model after training with CT learning using the Hebb rule ($Mdn = 1.06$), $U = 186.00,\ p < 0.001,\ r = 0.86$. Single cell information was also significantly lower in the untrained model compared to the model after training with CT learning combined with the trace rule ($Mdn = 1.48$), $U = 0.00,\ p < 0.001,\ r = 0.87$. Most significantly, single cell information was significantly higher when the model was trained using CT learning combined with the trace rule compared to CT learning alone, $U = 1551.00,\ p < 0.001,\ r = 0.84$.

In order to visualize the selectivity, we first identified the five cells that carried the highest single cell information regarding each identity of faces. We then recorded the firing rates of each of these cells in response to the presentation of all the four faces at all the 100 retinal locations. Fig. 5 shows the results in the untrained network (Fig. 5a), and networks trained with (Fig. 5b) CT learning, and CT learning with the trace rule (Fig. 5c). The figures presented on the top of each pane show the average firing rates of the five cells in response to the entire set of stimuli presented during testing, and the figures presented on the bottom show the mean firing rates for each face identity across the 100 transforms.

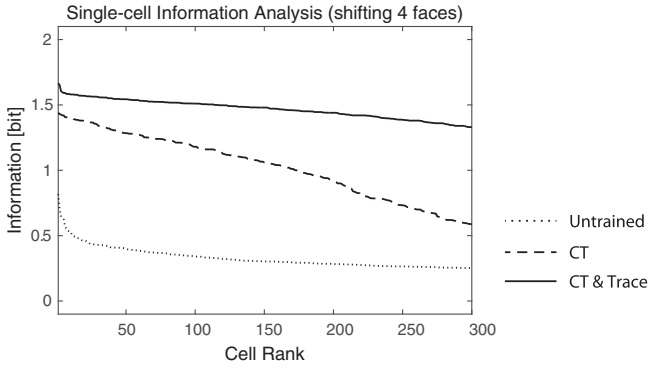

Fig. 4. Single cell information from experiment 1, using 4 faces in 100 locations. Results displayed are for the untrained network, CT learning using the Hebb rule (CT), and CT Learning combined with trace learning using the trace rule (CT and Trace). The plots show the maximum single cell information for 300 output cells plotted in rank order.

3.1.2. Model performance with 10 faces in 100 locations

In this simulation, we compared the performance of the model using the different learning rules when a large number of faces and a large number of locations were used (Fig. 6). In order to do this we tested the performance of the network with a stimulus set including 10 faces in 100 different locations ($N = 10,\ I_{max} = 3.32$).

We also found single cell information measures to be significantly lower in the untrained model ($Mdn = 0.47$) compared to the model after training with CT learning using the Hebb rule ($Mdn = 0.81$), $U = 1273.5,\ p < 0.001,\ r = 0.84$. Single cell information was also significantly lower in the untrained model compared to the model after training with CT learning combined with the trace rule ($Mdn = 2.89$), $U = 0.00,\ p < 0.001,\ r = 0.87$. Again, we found that combining CT learning with the trace rule during training led to significantly higher single cell information compared to training the model with CT learning alone, $U = 0.00,\ p < 0.001,\ r = 0.87$.

In order to visualize the selectivity, we first identified the top five cells that carried the highest single cell information regarding each stimulus. We then recorded the firing rates of these cells in response to the presentation of all 10 faces at all 100 retinal locations. Fig. 7 shows the results in the untrained network (Fig. 7a), and networks trained with (Fig. 7b) CT learning, and CT learning with the trace rule (Fig. 7c).

3.2. Experiment 2: Rotation invariance

3.2.1. Model performance with 4 faces from 100 views

In experiment 2, we compared the performance of the model when developing view invariance of faces from a large number of viewing angles (Fig. 8). To achieve this we tested the performance of the network when it was trained and tested with 4 faces from 100 viewing angles ($N = 4,\ I_{max} = 2$).

We again found that single cell information measures were significantly lower in the untrained model ($Mdn = 0.35$) compared to the model after training with CT learning using the Hebb rule ($Mdn = 1.41$), $U = 0.00,\ p < 0.001,\ r = -0.87$. Similarly, single cell information was significantly lower in the untrained model compared to the model after training with CT learning combined with the trace rule ($Mdn = 1.61$), $U = 0.00,\ p < 0.001,\ r = 0.87$. Most importantly, we also found that training the model with CT learning combined with the trace rule led to higher single cell information compared to CT learning alone, though the effect size was reduced, $U = 27012.50,\ p < 0.001,\ r = 0.35$.

To visualize the selectivity, we identified the five cells that carried the highest single cell information regarding the stimulus. We then recorded the firing rates of these cells in response to the presentation of all the four faces at all the 100 rotational views. Fig. 9 shows the results in the untrained network (Fig. 9a), and networks trained with (Fig. 9b) CT learning, and CT learning with the trace rule (Fig. 9c).

3.2.2. Model performance with 10 faces from 100 views

In this simulation, we compared the performance of the model using the different learning rules when a large number of faces and

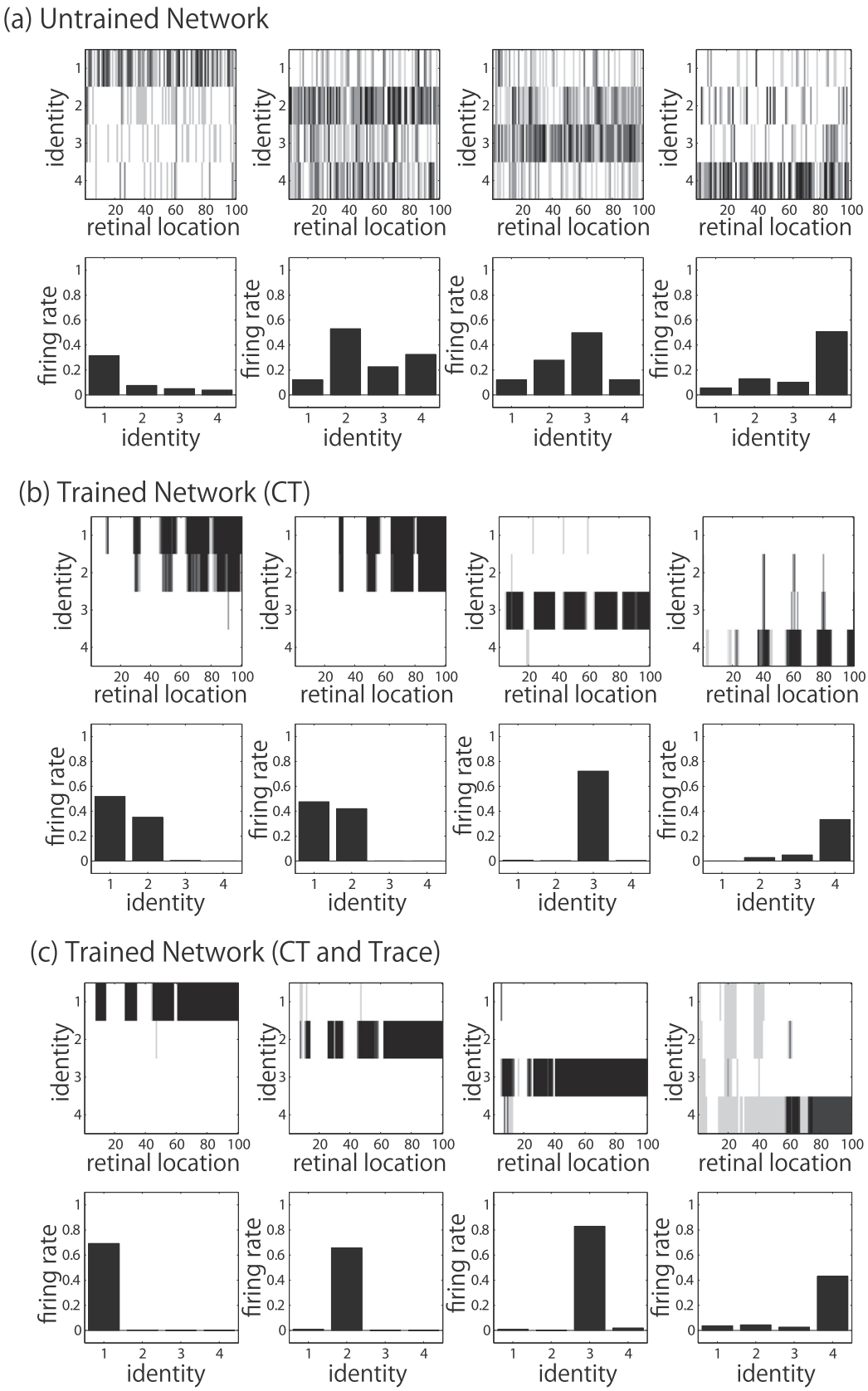

Fig. 5. Cell activations for four cells from experiment 1, using 4 faces in 100 locations. Plots show the cell selectivity based on the activation of the four most informative cells in the (a) untrained network, (b) trained network (CT), and (c) trained network (CT and Trace). For each stimulus, we first identified the top five cells that carried the highest single cell information regarding the stimulus. We then recorded the firing rates of these cells in response to the presentation of all four faces at all 100 retinal locations. The results are presented on the top of each pane, and the mean firing rates for each face identity across the 100 transforms are presented as a bar graph on the bottom of the pane.

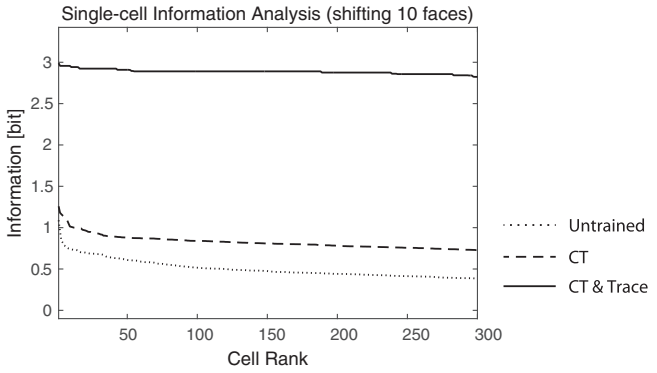

Fig. 6. Single cell information from experiment 1, using 10 faces in 100 locations. Results displayed are for the untrained network, CT learning using the Hebb rule (CT), and CT Learning combined with trace learning using the trace rule (CT and Trace). The plots show the maximum single cell information for 300 output cells plotted in rank order.

a large number of rotational views were used (Fig. 10). In order to do this we tested the performance of the network with a stimulus set including 10 faces in 100 different rotational views ($N = 10,\ I_{max} = 3.32$).

We also found single cell information measures to be significantly lower in the untrained model ($Mdn = 0.25$) compared to the model after training with CT learning using the Hebb rule ($Mdn = 0.94$), $U = 0,\ p < 0.001,\ r = 0.87$. Single cell information was also significantly lower in the untrained model compared to the model after training with CT learning combined with the trace rule ($Mdn = 1.31$), $U = 0.00,\ p < 0.001,\ r = 0.87$. Again, we found that combining CT learning with the trace rule during training led to significantly higher single cell information compared to training the model with CT learning alone, $U = 3305,\ p < 0.001,\ r = 0.80$.

To visualize the selectivity, we first identified the five cells that carried the highest single cell information regarding the stimulus. We then recorded the firing rates of these cells in response to the presentation of all 10 faces at all the 100 rotational views. Fig. 11 shows the results in the untrained network (Fig. 11a), and networks trained with (Fig. 11b) CT learning, and CT learning with the trace rule (Fig. 11c).

3.3. Experiment 3: Translation and rotation invariance

In experiment 3, we compared the performance of the model when developing both shifting and view invariance of faces from a large number of retinal locations and viewing angles (Fig. 12). To achieve this we tested the performance of the network when it was trained and tested with 4 faces from 50 viewing angles as well as 9 retinal locations ($N = 4,\ I_{max} = 2$). Each face at a particular viewing angle is shifted across different retinal locations in turn.

We again found that single cell information measures were significantly lower in the untrained model ($Mdn = 0.19$) compared to the model after training with CT learning using the Hebb rule ($Mdn = 0.65$), $U = 0.00,\ p < 0.001,\ r = 0.87$. Similarly, single cell

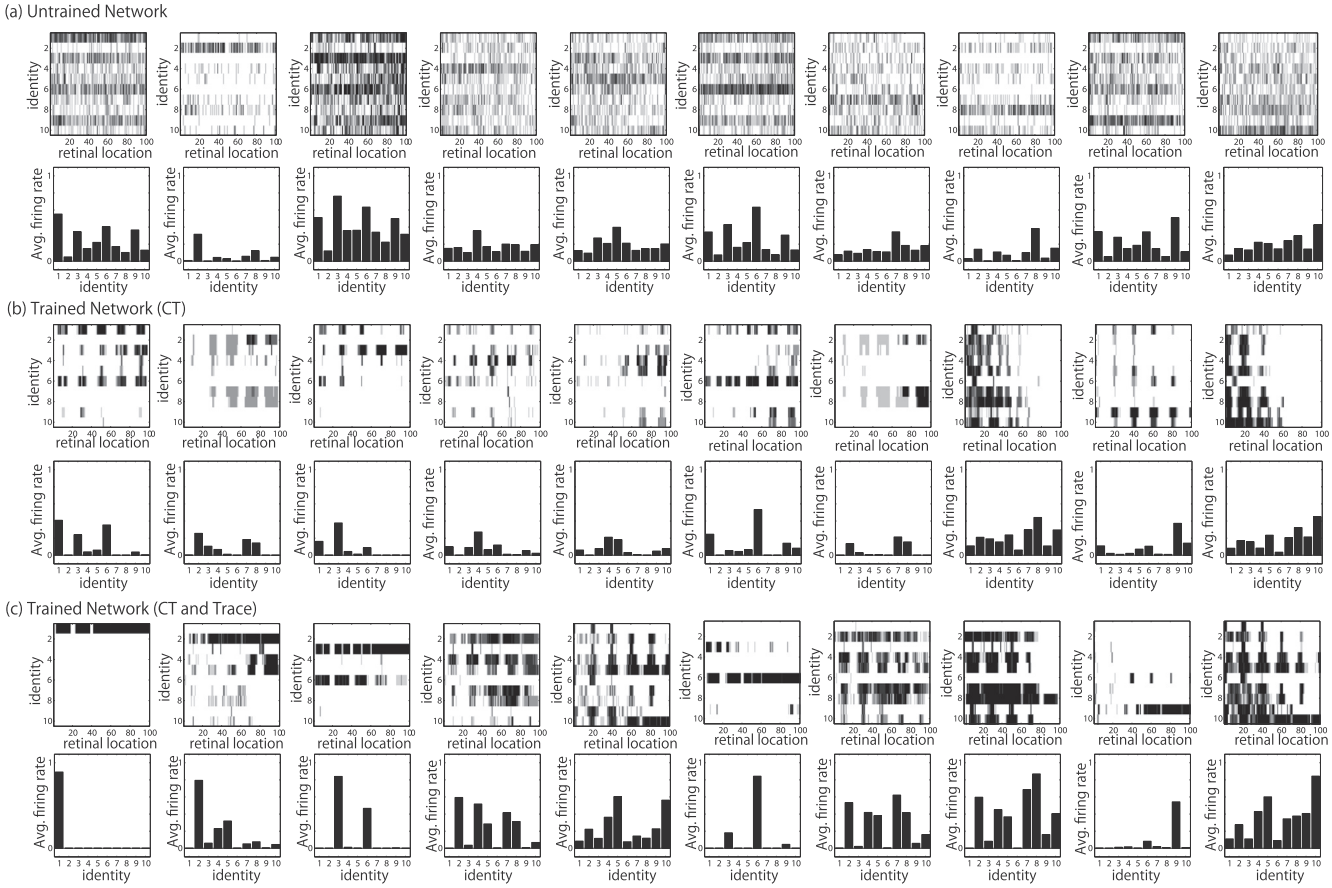

Fig. 7. Cell activations from experiment 1, using 10 faces in 100 locations. Plots show the cell selectivity based on the activations of subsets of cells in the (a) untrained network, (b) trained network (CT), and (c) trained network (CT and Trace). For each stimulus, we first identified the top five cells that carried the highest single cell information regarding the stimulus. We then recorded the firing rates of these cells in response to the presentation of all ten faces at all 100 retinal locations. The results are presented on the top of each pane, and the mean firing rates for each face identity across the 100 different views are presented as a bar graph on the bottom of the pane.

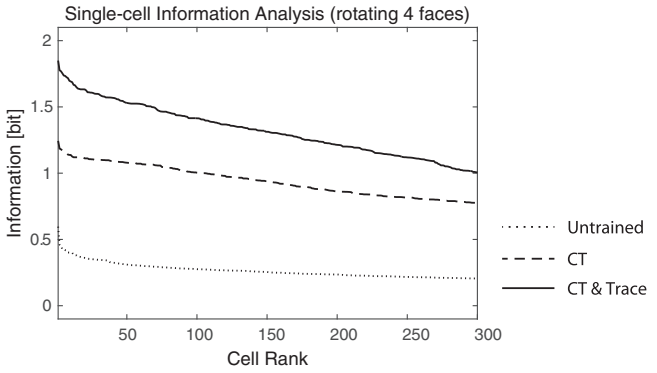

Fig. 8. Single cell information from experiment 2, using 4 faces from 100 viewing angles. Results displayed are for the untrained network, CT learning using the Hebb rule (CT), and CT Learning combined with trace learning using the trace rule (CT and Trace). The plots show the maximum single cell information for 300 output cells plotted in rank order.

information was significantly lower in the untrained model compared to the model after training with CT learning combined with the trace rule ($Mdn = 1.26$), $U = 0.00,\ p < 0.001,\ r = 0.87$. Most importantly, we also found that training the model with CT learning combined with the trace rule led to single cell information compared to CT learning alone, $U = 0.00,\ p < 0.001,\ r = 0.87$.

In order to visualize the selectivity, we first identified the five cells that carried the highest single cell information regarding the stimulus. We then recorded the firing rates of each subset of the cells in response to the presentation of all the four faces at all the 9 retinal locations and from 50 rotational views. Fig. 13 shows the results in the untrained network (Fig. 13a), and networks trained with (Fig. 13b) CT learning, and CT learning with the trace rule (Fig. 13c).

4. Discussion

The results shown here illustrate that using the trace rule in conjunction with CT learning can improve object-selective translation and view invariance in VisNet beyond what CT learning can achieve by itself. Furthermore, this effect is found to be present for both a small (4) and large (10) number of faces in simulations with location invariance and view invariance.

Inspecting the firing properties of individual cells in the model suggests that introducing the trace rule generally improves invariance by reducing the risk of cells learning to bind together the representations of different faces in a particular spatial location or viewing angle. (E.g., Figs. 5(b) and 7) This is because CT learning by itself has no information about the temporal sequence in which the images were presented. Therefore, as long as any two images appear similar in terms of spatial structure, the cells will learn to associate them together even if they are two different objects presented at different time points. As a result, the cells lose their face selectivity and respond to a large number of faces viewed in the same location or from the same angle.

As can be seen in our results, invariant representations of faces in the model worsen when the number of face identities increases. This is due to the increased difficulty of the task without a change in the capacity of the network (e.g. the size of the layers). If the difficulty of the task were to continue increasing in this manner, then the performance of the network would continue to decrease. However, even with this degraded overall performance, we still see a benefit of combining continuous transformation learning with temporal trace learning over using continuous transformation learning in isolation.

However, when the trace rule is used together with CT learning then temporal information also guides learning in the network. This temporal information can then be used to prevent the network from associating similar objects together that occur at widely different time points. Exploiting the temporal information means associations will not just be formed based on spatial similarity of the input, helping to increase the face selectivity in the network and strengthening invariance in the cells. This result is consistent with both psychophysical and neurophysiological data that suggest a role for temporal trace learning in developing and maintaining object-selective transform invariant representations of objects (Perry et al., 2006; Li & DiCarlo, 2008). This suggests that CT learning and temporal trace learning operating together could explain why invariant recognition of objects is improved when objects transform continuously in both space and time.

Whilst it is relatively straightforward to compare CT learning alone to CT learning combined with the trace rule, it is difficult to compare these two learning conditions with learning using the trace rule without CT learning. This is because the only way to prevent CT learning from taking place is to increase the spacing between stimulus locations. If the spacing is increased, but the number of training locations are kept the same, then the locations to be learnt will be spread over a larger retinal space, which is a more difficult task for the model. However, if the spacing is increased and the retinal space covered by the stimuli is kept the same, then there are fewer locations to develop invariance over, resulting in an easier task. This means that any comparison between CT learning and the trace rule without CT learning would be difficult to make.

4.1. Comparison to other computational modeling studies

Compared to the other work with convolutional networks (Krizhevsky, Sutskever, & Hinton, 2012; Taigman, Yang, Ranzato, & Wolf, 2014), one may assume that the problem of transform invariant representations of faces is tiny. However, the series of studies conducted in this paper represents an important theoretical advance in understanding how the visual system may learn such transform invariant representations, contrasting many current accounts of learning based on the feedback of error signals from higher- to lower-levels of representation.

In particular, our approach differs from the typical engineering approaches that rely on supervised signals in terms of the backpropagation of error (Rumelhart, Hinton, & Williams, 1986) to achieve accurate classifications of faces (Krizhevsky et al., 2012). In terms of neurophysiology, it is unlikely that the brain employs backpropagation of error because the mechanism requires finely structured neural connectivity, and such a network structure and the organizational principles required for its generation at the level of individual neurons is rather artificial and not biologically plausible (Stork, 1989).

Furthermore, for the development of transform invariant representations, many deep neural networks usually require large set of labeled data for invariances learning (Zou, Ng, & Yu, 2011); however, this training is not realistic as our brains are rarely provided with such ideal sets of visual inputs with labels. Although it has recently been reported that the accuracy of the transform invariant natural face verification by a deep neural network model has reached to nearly the same level of humans (Taigman et al., 2014), the algorithm still requires the employment explicit 3D face modeling to handle the transform invariance, which is again very biologically implausible.

Other attempts have been made to demonstrate invariant object recognition using teaching signals that are based on information that is inherent in the input, such as maximum likelihood cost functions (Becker, 1999). Although these models

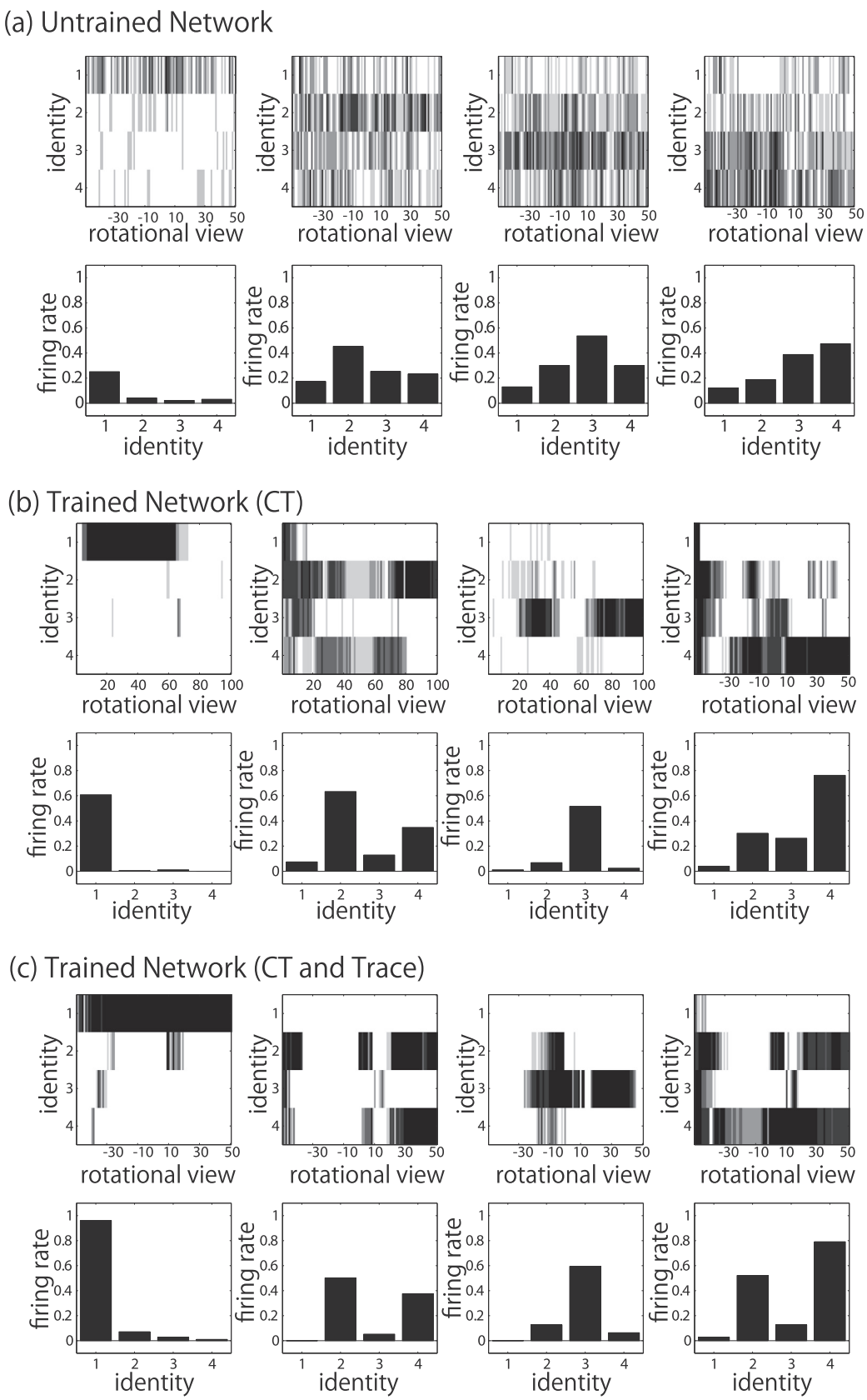

Fig. 9. Cell activations from experiment 2, using 4 faces from 100 viewing angles. Plots show cell selectivity based on the activations of subsets of cells in the (a) untrained network, (b) trained network (CT), and (c) trained network (CT and Trace). For each stimulus, we first identified five cells that carried the highest single cell information regarding the stimulus. We then recorded the firing rates of each subset of these cells in response to the presentation of all four faces at all 100 rotational views. The results are presented on the top of each pane, and the mean firing rates for each face identity across the 100 views are presented as a bar graph on the bottom of the pane.

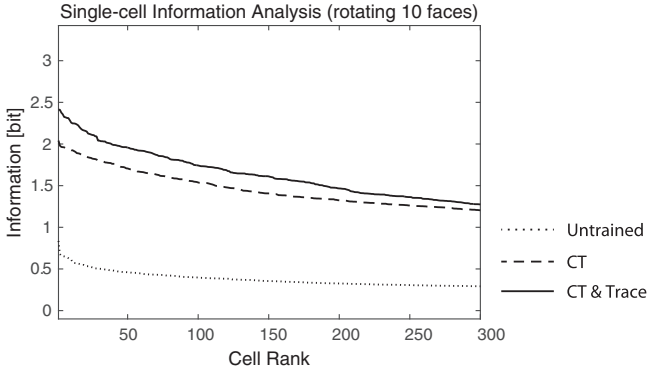

Fig. 10. Single cell information from experiment 2, using 10 faces from 100 viewing angles. Results displayed are for the untrained network, CT learning using the Hebb rule (CT), and CT Learning combined with trace learning using the trace rule (CT and Trace). The plots show the maximum single cell information for 300 output cells plotted in rank order.

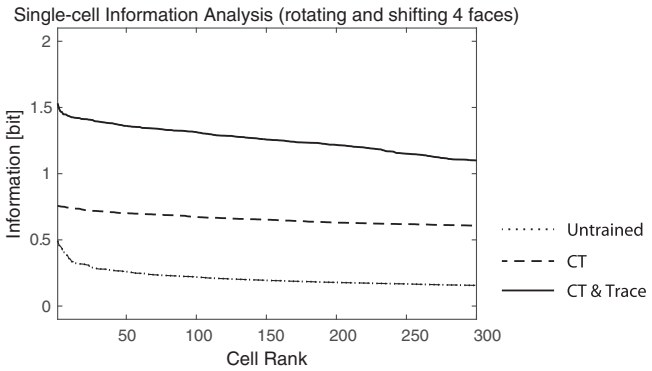

Fig. 12. Single cell information from experiment 3. Results displayed are for the untrained network, CT learning using the Hebb rule (CT), and CT Learning combined with trace learning using the trace rule (CT and Trace). The plots show the maximum single cell information for 300 output cells plotted in rank order.

are unsupervised, in the sense that the teaching signal is not based on external information, they still use backpropagation of error to learn the weights, which, as we have discussed above, is itself biologically implausible.

Besides the temporal trace learning rule (Földiák, 1991), Slow Feature Analysis (SFA) has been proposed as one of the few other unsupervised frameworks to solve the invariance problem in our brains (Berkes & Wiskott, 2005; Wiskott & Sejnowski, 2002). This is based on similar assumptions to temporal trace learning, as object identity tends to change much slower than the rapidly changing sensory input signals due to its transforms. Accordingly, they hypothesize that the neural systems in our brains are naturally self-organized in the way to detect a set of slowly changing features in the temporal sequence of visual inputs. By exploiting this principle, (Franzius, Wilbert, & Wiskott, 2008) has shown that the model can learn not only transform invariant object-specific representations but also the representations of other parameters

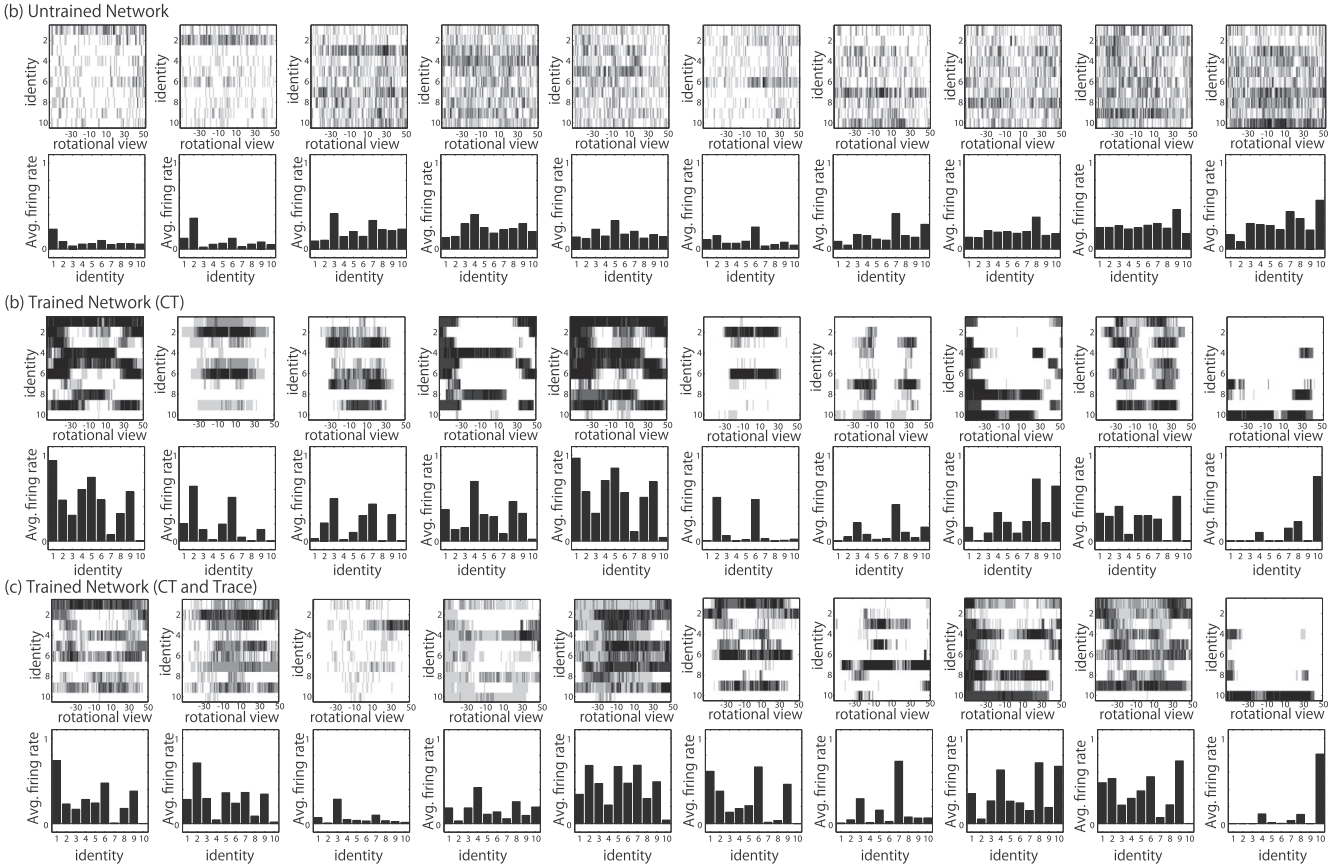

Fig. 11. Cell activations from experiment 2, using 10 faces from 100 viewing angles. Plots show the cell selectivity based on the activations of subsets of cells in the (a) untrained network, (b) trained network (CT), and (c) trained network (CT and Trace). For each stimulus, we first identified the top five cells that carried the highest single cell information regarding the stimulus. We then recorded the firing rates of each subset of these cells in response to the presentation of all ten faces at all 100 rotational views. The results are presented on the top of each pane, and the mean firing rates for each face identity across the 100 different views are presented as a bar graph on the bottom of the pane.

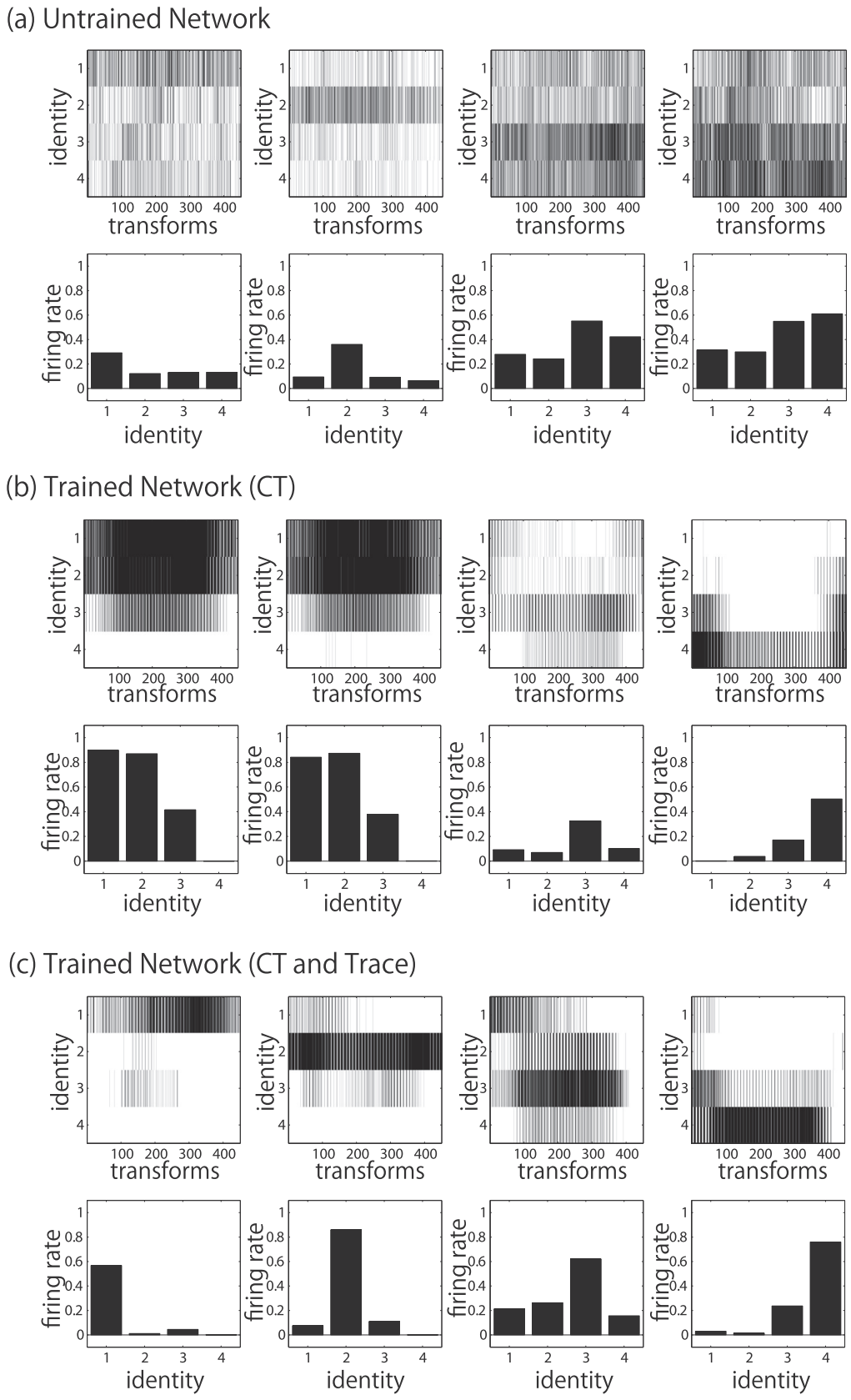

Fig. 13. Cell activations from experiment 3. Plots show cell selectivity based on the activations of subsets of cells in the (a) untrained network, (b) trained network (CT), and (c) trained network (CT and Trace). For each stimulus, we first identified the top five cells that carried the highest single cell information regarding the stimulus. We then recorded the firing rates of each subset of cells in response to the presentation of all four faces at all 9 retinal locations and at all 50 rotational views. The results are presented on the top of each pane, and the mean firing rates for each face identity across the 450 transforms are presented as a bar graph on the bottom of the pane.

Table A.5
Example cell firing rates to each face over presented in 100 different spatial locations.

Faces	$0 \leqslant r < 0.33$	$0.33 \leqslant r < 0.67$	$0.67 \leqslant r \leqslant 1$
A	3	17	80
B	68	31	1
C	73	25	2
D	65	12	17

such as object positions and rotation angles. However, in reality, SFA requires non-linear expansion of input, which suffers from the course of dimensionality (Zou et al., 2011). These mechanisms can be contrasted with our approaches where all computations required for the development of invariant representations in our model is achieved locally at each synapse.

Furthermore, based on the principle of SFA, (George & Hawkins, 2005) extended the algorithm to incorporate a Bayesian inference prediction framework and revised its architecture in accordance with the anatomic organization of the brain. As a result, this brain-inspired mechanism successfully exhibits invariance across a wide variety of transformations. Nevertheless, it is still not clear how exactly the brain may represent such information and develop the neuronal circuits to specifically extract slowly changing features.

Therefore, despite the relative simplicity of rate-coded models, such as VisNet, they still remain useful models for understanding competitive learning in vision. Their simplicity allows us to clarify the key learning mechanisms of interest and avoids additional confounding factors that might arise in more complex models, such as integrate-and-fire networks, justifying its current use.

In this paper, we present neural network simulations of the visually-guided development of transform invariant facial representations in the primate ventral visual pathway, using completely unsupervised learning mechanisms and feed-forward processing. In particular, VisNet differs from those other models in which the ability to utilize the temporal information is incorporated into the associative learning rules that govern the changes of synaptic weights. We believe this approach has implications for better learning rules in the current realm of deep learning by allowing networks to make use of additional information that might be inherent in the ordering of stimulus presentations. This may be particularly useful in many real world applications that involve sequences of inputs where deep learning can be used, such as processing speech and real-time visual inputs.

4.2. Conclusions

Previous research has shown that invariant representations of objects can be learnt based on temporal information in VisNet (Wallis & Rolls, 1997) and in other models of invariant object recognition (Becker, 1999; George & Hawkins, 2005). Furthermore, we have known that CT learning on its own can produce invariant representations of objects (Stringer et al., 2006). However, this research has demonstrated for the first time that trace learning and CT learning can work cooperatively to improve invariance learning in VisNet. More generally, this suggests that neural network architectures that perform transform invariant object recognition based on structural similarity could benefit from also making use of temporal information.

More specifically, we have shown that, whilst some invariance can be achieved based upon CT learning alone, the visual system can use temporal trace learning to prevent cells losing their selectivity to particular objects. This is achieved by providing the system with temporal information about object translation, allowing both mechanisms to cooperate within the same model. We know that this type of trace learning naturally emerges in biophysically realistic spiking neural networks (Evans & Stringer, 2012). Therefore, future research may wish to explore whether cooperation between CT learning and temporal trace learning can emerge in spiking neural networks. Furthermore, this research highlights the fact that the visual system is unlikely to use single cues to develop invariant object representations, but is likely to use cues from multiple sources during learning - such as spatial and temporal information – in order to achieve this feat.

Appendix A. Computing cell information measures

The single cell information measure used in these simulations is given by Eq. (10). The process of computing these values will be given in more detail here by the use of an example.

In this case we will consider single cell information measures for simulations with four different faces, A, B, C and D, and 100 different spatial locations. As each face is presented an equal number of times the probability of each face being presented, $P(s)$ will be $P(s) = 1/4$. To calculate the probability of each response the firing rates for each cell, r, must be binned. We chose to use three equally spaced bins, $0 \leqslant r < 0.33$, $0.33 \leqslant r < 0.67$, and $0.67 \leqslant r \leqslant 1$. This produces a matrix of responses for each cell, an example is given in Table A.5.

Using the table of firing rates we can calculate the information that a particular response from the cell carries about a particular stimulus by calculating the probability of that response $P(r)$ and the probability of the responses given the stimulus $P(r|s)$. For example, the strongest category of response $0.67 \leqslant r \leqslant 1$ has the probability of occurring $P(r) = 100/400 = 0.25$ and the probability of occurring given that face A was presented of $P(r|s) = 80/100 = 0.8$. Therefore, by Eq. (10) the amount of information about face A carried by this category of response is $I(s, R) = 0.8 \log_2 0.8/0.25 = 0.931$.

The information value given for each cell is the maximum conveyed by a particular response about a particular stimulus. In the case of this example, the information for this cell would be given as 0.931.

References

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

Berkes, P., & Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision, 5*(6), 9. http://dx.doi.org/10.1167/5.6.9. ISSN 1534-736.

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

Cumming, B. G., & Parker, A. J. (1999). Binocular neurons in V2 of awake monkeys. *The Journal of Neuroscience, 16*, 5602–5618.

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

Destexhe, A., & Pare, D. (1999). Impact of network activity on the integrative properties of neocortical pyramidal neurons in vivo. *Journal of Neurophysiology, 81*(4), 1531–1547. ISSN 0022-3077.

Eguchi, A., Neymotin, S. A., & Stringer, S. M. (2014). Color opponent receptive fields self-organize in a biophysical model of visual cortex via spike-timing dependent plasticity. *Frontiers in Neural Circuits, 8*(16). http://dx.doi.org/10.3389/fncir.2014.00016.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., & Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research, 11*, 625–660. ISSN 1532-4435.

Evans, B. D., & Stringer, S. M. (2012). Transformation-invariant visual representations in self-organizing spiking neural networks. *Frontiers in Computational Neuroscience, 6*.

Evans, B. D., & Stringer, S. M. (2013). How lateral connections and spiking dynamics may separate multiple objects moving together. *PLoS ONE, 8*(8), e69952. http://dx.doi.org/10.1371/journal.pone.0069952.

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

Franzius, M., Wilbert, N., & Wiskott, L. (2008). Invariant object recognition with slow feature analysis. In V. Kurková, R. Neruda, & J. Koutnk (Eds.), *Artificial neural networks – ICANN 2008. No. 5163 in lecture notes in computer science* (pp. 961–970). Berlin Heidelberg: Springer. ISBN 978-3-540-87535-2 978-3-540-87536-9.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience, 14*, 1195–1201.
George, D., & Hawkins, J. (2005). A hierarchical Bayesian model of invariant pattern recognition in visual cortex. *IEEE international joint conference on neural networks* (vol. 3, pp. 1812–1817). IEEE.
Hasselemo, M. E., Rolls, E. T., Baylis, G. C., & Nalwa, V. (1989). Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research, 75*, 415–429.
Ito, M., Tamura, H., Fujita, I., & Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology, 73*, 218–226.
Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in the cat striate cortex. *Journal of Neurophysiology, 58*, 1233–1258.
Kobotake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology, 71*, 856–867.
Kohonen, T. (1982). Cluseter, taxonomy, and topological maps of patterns. In M. Lang (Ed.), *Proceedings of the sixth international conference on pattern recognition* (pp. 114–125). MD: IEEE Computer Society Press, Silver Spring.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems* (vol. 25, pp. 1097–1105). Curran Associates Inc.
Li, N., & DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science, 321*, 1502–1507.
Mainen, Z., & Sejnowski, T. (1995). Reliability of spike timing in neocortical neurons. *Science, 268*(5216), 1503–1506. ISSN 0036-8075.
Marreiros, A. C., Daunizeau, J., Kiebel, S. J., & Friston, K. J. (2008). Population dynamics: Variance and the sigmoid activation function. *NeuroImage, 42*(1), 147–157. http://dx.doi.org/10.1016/j.neuroimage.2008.04.239. ISSN 1053-8119.
Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., Le, Q.V., Ng, A.Y. (2011). On optimization methods for deep learning. In: *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 265–272).
Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology, 14*(4), 481–487. http://dx.doi.org/10.1016/j.conb.2004.07.007. ISSN 0959-4388.
Op de Beeck, H., & Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. *Journal of Comparative Neurology, 426*, 505–518.
Pasupathy, A. (2006). Neural basis of shape representation in the primate brain. *Progress in Brain Research, 154*, 293–313.
Perry, G., Rolls, E. T., & Stringer, S. M. (2006). Spatial vs temporal continuity in view invariant visual object recognition learning. *Vision Research, 46*, 3994–4006.
Petkov, N., & Kruizinga, P. (1997). Computational models of visual neurons specialised in the detection of periodic and aperiodic oriented visual stimuli: Bar and grating cells. *Biological Cybernetics, 76*, 83–96.
Pettet, M. W., & Gilbert, C. D. (1992). Dynamic changes in receptive-fields size in cat primary visual cortex. *Proceedings of the National Academy of Sciences, 89*, 8366–8370.
Ranzato, M., Huang, F. J., Boureau, Y.-L., LeCun, Y. (2007). Unsupervised learning of invariant feature hierarchies with applications to object recognition, IEEE, 1–8. http://dx.doi.org/10.1109/CVPR.2007.383157. ISBN 978-1-4244-1179-5 978-1-4244-1180-1.
Rolls, E. T. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philosophical Transactions of the Royal Society, 335*, 11–21.
Rolls, E. T. (2000). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron, 27*, 205–218.
Rolls, E. (2007). *Memory, attention, and decision-making: A unifying computational neuroscience approach*. 1st ed.: Oxford University Press. ISBN 978-0-19-923270-3.
Rolls, E. T., & Baylis, G. C. (1986). Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Experimental Brain Research, 65*, 38–48.
Rolls, E. T., Baylis, G. C., & Hasselmo, M. E. (1987). The responses of neurons in the cortex of the superior temporal sulcus of the monkey to bandpass spatial frequency filtered faces. *Vision Research, 27*, 311–326.
Rolls, E. T., Baylis, G. C., & Leonard, C. M. (1985). Role of low and high spatial frequencies in the face-selective responses of neurons in the cortex in the superior temporal sulcus. *Vision Research, 25*, 1021–1035.
Rolls, E. T., & Deco, G. (2002). *Computational neuroscience of vision*. Oxford: Oxford University Press.
Rolls, E. T., & Milward, T. (2000). A model of invariant object recognition in the visual system: Learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Computation, 12*, 2547–2572.
Rolls, E., & Tovee, M. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology, 73*(2), 713–726. ISSN 0022-3077.
Rolls, E. T., & Treves, A. (1998). *Neural networks and brain function*. Oxford: Oxford University Press.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536. http://dx.doi.org/10.1038/323533a0.
Stork, D. (1989). Is backpropagation biologically plausible? In: *International joint conference on neural networks, 1989. IJCNN, vol. 2* (pp. 241–246). http://dx.doi.org/10.1109/IJCNN.1989.118705.
Stringer, S. M., Perry, G., Rolls, E. T., & Proske, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biological Cybernetics, 94*, 128–142.
Taigman, Y., Yang, M., Ranzato, M., Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. In: *2014 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1701–1708). http://dx.doi.org/10.1109/CVPR.2014.220.
Tanaka, K., Saito, H., Fukada, Y., & Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology, 66*, 170–189.
Tovee, M. J., Rolls, E. T., & Azzopardi, P. (1994). Translation invariance and the responses of neurons in the temporal visual cortical areas of primates. *Journal of Neurophysiology, 72*, 1049–1060.
Tromans, J. M., Harris, M., & Stringer, S. M. (2011). A computational model of the development of separate representations of facial identity and expression in the primate visual system. *PLoS ONE, 6*(10), e25616. http://dx.doi.org/10.1371/journal.pone.0025616.
Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science, 287*(5456), 1273–1276. http://dx.doi.org/10.1126/science.287.5456.1273. ISSN 0036-8075, 1095-920.
Vogels, R., & Biederman (2002). Effects of illumination intensity and direction on object coding in the macaque inferior temporal cortex. *Cerebral Cortex, 12*, 756–766.
Von der Marlsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik, 14*, 85–100.
Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology, 51*, 167–194.
Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation, 14*(4), 715–770. http://dx.doi.org/10.1162/089976602317318938. ISSN 0899-766.
Zou, W. Y., Ng, A. Y., Yu, K. (2011). Unsupervised learning of visual invariance with temporal coherence. In: *NIPS 2011 workshop on deep learning and unsupervised feature learning*.