NETWORK: COMPUTATION IN NEURAL SYSTEMS
2016, VOL. 27, NO. 1, 29–51
http://dx.doi.org/10.1080/0954898X.2016.1187311

Taylor & Francis
Taylor & Francis Group

The visual development of hand-centered receptive fields in a neural network model of the primate visual system trained with experimentally recorded human gaze changes

Juan M. Galeazzi[a], Joaquín Navajas[b,c], Bedeho M. W. Mender[a], Rodrigo Quian Quiroga[c], Loredana Minini[a], and Simon M. Stringer[a]

[a]Oxford Centre for Theoretical Neuroscience and Artificial Intelligence, Department of Experimental Psychology, University of Oxford, Oxford, UK; [b]Institute of Cognitive Neuroscience, University College London, London, UK; [c]Centre for Systems Neuroscience, University of Leicester, Leicester, UK

ABSTRACT

Neurons have been found in the primate brain that respond to objects in specific locations in hand-centered coordinates. A key theoretical challenge is to explain how such hand-centered neuronal responses may develop through visual experience. In this paper we show how hand-centered visual receptive fields can develop using an artificial neural network model, VisNet, of the primate visual system when driven by gaze changes recorded from human test subjects as they completed a jigsaw. A camera mounted on the head captured images of the hand and jigsaw, while eye movements were recorded using an eye-tracking device. This combination of data allowed us to reconstruct the retinal images seen as humans undertook the jigsaw task. These retinal images were then fed into the neural network model during self-organization of its synaptic connectivity using a biologically plausible *trace learning* rule. A trace learning mechanism encourages neurons in the model to learn to respond to input images that tend to occur in close temporal proximity. In the data recorded from human subjects, we found that the participant's gaze often shifted through a sequence of locations around a fixed spatial configuration of the hand and one of the jigsaw pieces. In this case, trace learning should bind these retinal images together onto the same subset of output neurons. The simulation results consequently confirmed that some cells learned to respond selectively to the hand and a jigsaw piece in a fixed spatial configuration across different retinal views.

ARTICLE HISTORY
Received 7 January 2016
Revised 4 May 2016
Accepted 4 May 2016

KEYWORDS
Eye-tracker; hand-centered; neural networks; reference frames; VisNet; trace learning

1. Introduction

Different regions of the visuomotor pathway in the primate brain contain neurons that represent the locations of visual targets in different nonretinal coordinate frames linked to different parts of the body. Several

CONTACT Juan M. Galeazzi juan.galeazzigonzalez@psy.ox.ac.uk Department of Experimental Psychology, University of Oxford, Tinbergen Building, 9 South Parks Road, Oxford OX1 3UD, UK.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/inet.

neurophysiological studies have reported cells in different parts of the posterior parietal cortex (PPC) and adjacent premotor areas encoding the location of visual targets in a continuum of coordinate frames that are anchored to the hand, head and body/trunk (Andersen et al. 1985, Bremner and Andersen 2012, Brotchie et al. 1995, Buneo et al. 2002, Pesaran et al. 2006). These representations are thought to be used to guide motor plans linked to their corresponding effectors.

For example, cells in area 5d are putatively involved in representing the location of visual targets in hand-centered coordinates (Bremner and Andersen 2012, Buneo and Andersen 2006). Hand-centered receptive fields or peri-hand representations have also been reported in the ventral premotor area (PMv) and other areas (Graziano et al. 1994, 1997, Graziano and Gross 1998). These cells fire maximally whenever the target is in a preferred location relative to the hand. Furthermore, in many cases these neuronal responses seem to be invariant to shifts in retinal position due to gaze changes effected by eye and head movements as well as by movements of the hand itself (Bremner and Andersen 2012). Similarly, studies with human primates have also shown evidence of hand-centered representations in the PPC and premotor areas (Makin et al. 2009, Brozzoli et al. 2012, 2011, Makin et al. 2007, Gentile et al. 2011). These hand-centered representations are thought to play a role in visually guided reaching to target objects, as well as potentially providing a mechanisms for avoidance reactions.

We have previously proposed a computational hypothesis of how *trace learning* may allow neurons to develop selective responses to the location of visual objects relative to the hand that are invariant to shifts in retinal position (Galeazzi et al. 2013). Trace learning is a biologically plausible learning mechanism that encourages cells to learn to respond to input images that tend to occur in close temporal proximity (Földiák 1991). This is achieved by incorporating a memory trace of the recent neuronal activity into a local associative learning rule. We proposed that, for a portion of the time, humans shift their eyes around static visual scenes that contain their hand with other nearby objects in a fixed spatial configuration. In this case, trace learning will bind together these retinal images onto the same subset of higher layer neurons, which will then respond to particular hand-object configurations regardless of retinal position. Such cells effectively encode the hand-centered locations of visual targets, as reported in neurophysiology studies (Bremner and Andersen 2012). This hypothesis was tested in our unsupervised, self-organizing neural network model, VisNet, of the primate visual system. Our simulations confirmed the plausibility of this hypothesis, and showed how different output cells learned to respond selectively to different object positions relative to the hand (Galeazzi et al. 2013). More recently, we have demonstrated the ability of our model to develop hand-centered visual representations even when it is trained using highly realistic

images, in which the hand is seen against natural scenes with multiple objects present at the same time (Galeazzi et al. 2015).

However, despite the recent improvements in the realism of the images on which VisNet was successfully trained, the dynamics of the eye movements were still unrealistic and controlled artificially. The simulations in Galeazzi et al. (2013, 2015) used only a limited number of equidistant, prespecified shifts (five or six retinal shifts in total) during training and testing. The richness and complexity of the dynamics of natural eye movements from human test subjects has never been explicitly incorporated to guide the retinal shifts in VisNet during training. More importantly, by substantially increasing the number of retinal shifts during training, the associative (Hebbian) component of the trace learning rule could have unwanted deleterious effects. For example, smooth and continuous retinal shifts could generate substantial spatial overlap between some of the images fed to the network during training. A continuous transformation (CT) learning mechanism (Stringer et al. 2006) binds together spatially overlapping visual stimuli. This could enable CT learning to bind together different hand-centered locations by the same cell and therefore seriously degrade the hand-centered location specificity of neurons.

Furthermore, previous research with VisNet has mainly represented time in discrete processing steps, in which a time step corresponds to an unspecified interval of time. However, in order to feed video images to the network that faithfully represent the temporal dynamics of gaze changes recorded from participants, we needed to implement a new time accurate differential formulation of the VisNet model. It is also important to explicitly define the dynamical quantities and parameters that would govern the network, given that we are assuming that a temporal trace in the neuronal dynamics is what enables the binding of temporally concurrent views of the same hand-centered configuration. For example, in a differential version of trace learning, the trace value would be exponentially decaying through time while the neuron is not active. Therefore, these simulations would help us to establish whether trace learning could cope effectively and allow the network to form representations of hand-centered configurations which are invariant across retinal shifts, based on how we explore a visual scene in a hand-object manipulation task.

In this paper, we show how hand-centered visual receptive fields can develop during visually guided learning in VisNet when the model is driven by gaze changes recorded from human test subjects. The purpose of this study is to show how the statistics of natural eye and head movements are capable of driving the development of such hand-centered neuronal responses. Human participants undertook an experimental task in which they had to complete a jigsaw and were free to move their eyes and head. A camera mounted on the head captured images of the hand and jigsaw,

while eye movements were recorded using a head-mounted eye-tracking device. This combination of data allowed us to reconstruct the retinal images as they undertook the jigsaw task. These sequences of retinal images were then fed into the simulations of the neural network model during self-organization of its synaptic connectivity using trace learning.

We assumed that while a subject is solving a jigsaw puzzle, there would be periods of time in which the hand would remain stationary while the participants explored the visual scene composed of the hand and an object near the hand (i.e. a jigsaw puzzle piece). It has been previously shown that trace learning is robust enough to allow a competitive network to develop nonretinal representations even if the training regime is composed of a wide variety of stimuli dynamics, as long as the stimuli dynamics required for our self-organization hypothesis occur for a small portion of the time (Mender and Stringer 2014). Therefore, trace learning can still function even if the majority of the time, the dynamics of gaze changes and hand movements are different to what is required in our self-organizing hypothesis.

In the data recorded from the human participants, we used the periods of time in which their gaze often shifted through a sequence of locations around a fixed spatial configuration of the hand and a jigsaw piece to train the network. Trace learning in our model simulations was then able to bind these retinal images together onto the same subset of output neurons. In this way, trace learning encouraged cells to respond to particular hand-object configurations across different retinal locations. Thus, after training, some output cells responded to specific hand-centered locations irrespective of retinal position.

The simulations reported in this paper therefore provide a plausible model of how the retinal image sequences arising from natural movements of the eyes and head recorded from human participants are able to drive the development of the kind of hand-centered visual representations that have been found experimentally in the primate brain during single unit recording studies (Bremner and Andersen 2012, Buneo and Andersen 2006).

2. Materials and methods

2.1. *Experimental task*

2.1.1. *Participants*

Six adults (2 males, 4 females, mean age 22.3 years, SD 2.6 years) participated in this study. All of them were naïve and corrected-to-normal vision. One of the females was excluded from the experiment during the calibration phase due to a lack of a reliable signal from the pupil to the eye tracking equipment. This study will show the results from the five remaining participants.

All participants gave written informed consent approved by the ethics committee at the Centre for Systems Neuroscience, University of Leicester. This study was conducted in accordance with the Declaration of Helsinki (Code of Ethics of the World Medical Association).

2.1.2. *Apparatus*

Eye movements were recorded with the mobile eye tracker ASL MobileEye (Applied Science Laboratories; Bedford, MA, USA) at a sampling rate of 30 Hz. This video-based eye tracker is mounted on a pair of lightweight safety glasses containing two cameras. One of the cameras was directed to the right eye, which was tracked using the pupil center and corneal reflection. The second camera recorded the scene. We aligned the scene camera with the participant's line-of-sight, and captured a video image of their approximate visual field. The visual range of the scene camera was approximately 50 degrees horizontal and 40 degrees vertical. The eye position recorded by the eye-tracking camera was used to compute gaze direction within each scene frame. By combining the image taken by the scene camera with the gaze direction within the scene frame, we were able to reconstruct the participant's retinal image of the scene, which we then fed directly into the VisNet simulations.

2.1.3. *Materials*

In this study, we used a jigsaw puzzle manufactured by the UK company Gibsons©; the puzzle was made of good-quality cardboard, with 100 pieces. The puzzle displayed a painting from Trevor Mitchell called 'A Hop, Skip and a Jump'. The completed puzzle measured 49cm x 34cm (approx 19 x 13 inches). Participants did not have any prior experience with the puzzle.

2.1.4. *Procedure*

Participants sat on a chair in front of a table. They were informed that they would be completing a jigsaw puzzle that was on the table covered by a cloth and were asked to adjust their posture and distance from the table according to their own preference. At that moment we adjusted the camera recording the scene and aligned it with the forward line-of-sight of the participant and captured a video image of their approximate visual field. The scene camera was calibrated with the pupil recording in order to be able to track the gaze direction within the scene frames. The scene camera calibration required subjects to fixate on each of the four fixation crosses that were marked across the table. After calibration, subjects were asked to re-fixate in the previously instructed points to corroborate the calibration accuracy.

After the eye-tracking calibration was successful, the participants were informed about the details of the task. They were told that after the cue provided by the experimenters, the cloth covering the jigsaw puzzle would be

removed. Subjects were never shown the full picture that was depicted in the puzzle. The jigsaw was presented almost completed with only five missing pieces that had been previously removed by the experimenters. All of the pieces removed had exactly the same shape in order to make sure that subjects would not use the shape of the piece as a strategy to find the location of the missing piece.

Participants were instructed that the cloth covering the incomplete jigsaw would be removed signaling the beginning of the trial and one of the missing pieces would be placed near their hands. They would then need to look around and make a single decision to place this piece in its corresponding location once they were sure they had found it. There were no incorrect trials, and all participants managed to accurately find the corresponding location of the jigsaw piece. The puzzle piece could appear in one of three possible locations near the hand which was resting on the table in front them. These hand-centered locations were left of the hand, right of the hand, and in front of the hand. After the participants placed the piece in the corresponding puzzle location the trial was terminated. For the next trial the experimenters would remove five different pieces from the jigsaw puzzle and repeat the procedure. The presentation of the trials was counterbalanced across subjects, as well as the order in which the missing pieces were placed around the hand (i.e. left, right, in front).

Typically the participants would explore the visual scene and saccade for a few seconds between the hand and jigsaw piece and the possible locations in the puzzle. The average length of the trials varied between subjects, having an average of 10.09s with a standard deviation of 6.7s. In Figure 1, we can see an example of the gaze changes in one of the trials, with the jigsaw piece placed to the right of the hand. It can be seen that the human participant made a series of saccades between the hand, jigsaw piece and the puzzle. We used software to identify and record the locations of visual fixations within the scene, which are shown as dots in Figure 1.

2.2. *The VisNet architecture and model equations: A new time-accurate differential version of the model*

Previous research with VisNet has represented time in discrete processing steps, in which a time step corresponds to an unspecified interval of time. However, in order to feed video images to the network that faithfully represent the temporal dynamics of gaze changes recorded from human participants, we required a new time accurate differential formulation of the VisNet model. This section provides a description of the architecture and equations that govern the time accurate version of VisNet used for these simulations. Most of the architectural features of the model are similar to the previous discrete version of VisNet, which is described in Wallis and Rolls (1997), Rolls and Milward (2000), and

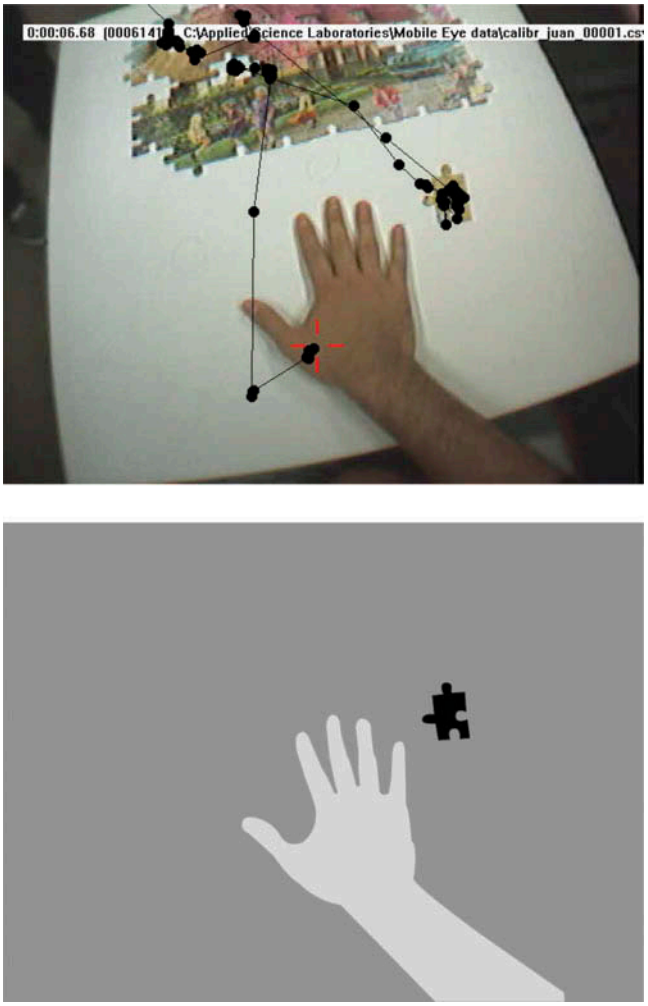

Figure 1. Example of the gaze changes during one of the trials and a sample of preprocessed frame. The upper image shows the dots that correspond to the locations of visual fixations extracted from numerical data provided by the software. In this trial the head remained stationary. Across participants, the eye movements typically involved saccades back and forth between the puzzle piece and the potential puzzle locations in which the piece could fit. The lower image shows an example preprocessed frame before it is presented to the network as per standard VisNet preprocessing procedures.

Rolls (2008). The time accurate version of VisNet also consists of four competitive layers of neurons with topologically organized feedforward synaptic connections between successive layers, as shown in Figure 2.

Even though VisNet was first conceived as a model of the primate ventral visual system, it has been more recently shown that the model is sufficiently robust to be applied to visual processes occurring in the dorsal system (Rolls and Stringer 2007, Rolls and Webb 2014), including the visual development of hand-centered receptive fields (Galeazzi et al. 2013, 2015).

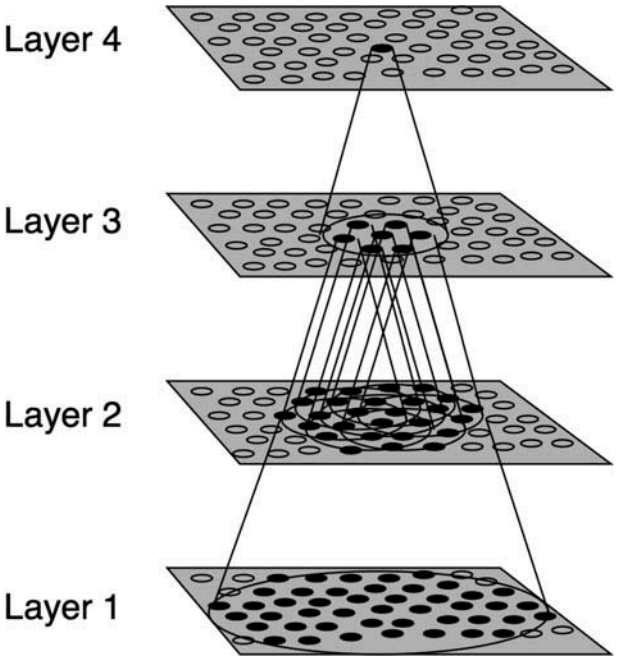

Figure 2. The VisNet model. VisNet neural network architecture. There are four layers of neurons with feedforward synaptic connections between successive layers that are hierarchically organized. The strengths of the feedforward connections are modified using a trace learning rule as the network is trained on images of the hand with a jigsaw piece in different locations around the hand.

The input images are first preprocessed by a layer of Gabor filters that mimic the responses of bar and edge-detecting simple cells in cortical visual area V1. The implementation of this filtering procedure has been described previously in Galeazzi et al. (2013). The x, y locations on the retina contain a bank of Gabor filter outputs tuned to different orientations and frequencies corresponding to a hypercolumn. Each of the Gabor filters is convolved with its corresponding local region of the image. The Gabor filters are given by

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

$$x' = x\cos\theta + y\sin\theta, \tag{2}$$

$$y' = -x\sin\theta + y\cos\theta \tag{3}$$

for all combinations of $\lambda = 2, \gamma = 0.5, \sigma = 0.56\lambda, \theta \in \{0, \pi/4, \pi/2, 3\pi/4\}$ and $\psi \in \{0, \pi, \pi/2, -\pi/2\}$. The outputs of the Gabor filters after convolution with the image are passed to the first layer of neurons in the network.

The neuronal activations and firing rates within each of the four layers is modeled as follows. Let $h_i(t)$ and $y_i(t)$ denote the activation and firing rate

respectively of neuron i in layer ℓ at time t, and let $w_{ij}(t)$ be the weight of the synapse from neuron j in layer $\ell - 1$ to neuron i in layer ℓ. The activation of each cell i is calculated according to

$$\tau_h \frac{dh_i}{dt} = -h_i(t) + \sum_j w_{ij}(t) y_j(t) \tag{4}$$

where τ_h is an activation time constant which is constant for all cells. The $-h_i(t)$ models the exponential decay of the activation $h_i(t)$ in the absence of any external input. In the simulations reported below, the activation time constant τ_h was set to relatively small values of 10 to 20 ms through successive layers of the network in order to ensure the neurons responded rapidly to the changing visual images. Equation (4) is solved numerically using a Forward Euler finite difference method.

The competition was implemented as previously described in VisNet studies (e.g. Galeazzi et al. (2013). Competition in each of the layers is graded and in this case is carried out in two stages at each numerical time step. In order to implement lateral inhibition within a layer, the activation of neurons are convolved using a spatial filter I, where δ regulates the contrast and σ determines the width, and a and b index the distance away from the centre of the filter

$$I_{a,b} = \begin{cases} -\delta e^{-\frac{a^2+b^2}{\sigma^2}} & \text{if } a \neq 0 \text{ or } b \neq 0, \\ 1 - \sum_{\substack{a \neq 0 \\ b \neq 0}} I_{a,b} & \text{if } a = 0 \text{ and } b = 0. \end{cases} \tag{5}$$

Afterwards, we apply contrast enhancement by means of a sigmoid activation function

$$y = f^{sigmoid}(r) = \frac{1}{1 + e^{-2\beta(r-\alpha)}}, \tag{6}$$

where r is the activation after lateral inhibition, y is the firing rate after contrast enhancement, and α and β are the sigmoid threshold and slope, respectively. The parameters α and β are constant within each layer, although α is adjusted to control the sparseness of the firing rates. For the simplified case of neurons with binarized firing rates $= 0/1$, the sparseness is the proportion $\in [0, 1]$ of neurons that are active. For example, to set the sparseness to 5% the threshold is set to the value of the 95th percentile point of the activations within the layer.

2.3. *Trace learning*

Different views of the hand and a jigsaw piece in a fixed spatial configuration will tend to occur close together in time as the eyes quickly saccade around the scene. In order to bind together these temporally concurrent input patterns during learning we used a trace learning rule that incorporated a

memory trace of recent neuronal activity. The trace value $\overline{y}_i$ for neuron i is governed by

$$\tau_{\overline{y}} \frac{d\overline{y}_i}{dt} = -\overline{y}_i + y_i, \tag{7}$$

where $\tau_{\overline{y}}$ is a trace time constant that is common for all the cells. In the simulations shown below, $\tau_{\overline{y}}$ was set to a relatively large value of 500ms in order to allow postsynaptic neurons to bind together temporally contiguous input patterns corresponding to different retinal locations. Moreover, additional simulations demonstrated robust performance across the range $\tau_{\overline{y}} \in [100ms, 500ms]$.

We then update the strength of the synaptic weight from presynaptic neuron j to postsynaptic neuron i using a differential form of the trace learning rule

$$\frac{dw_{ij}}{dt} = \alpha \overline{y}_i y_j, \tag{8}$$

where α is the learning rate, y_j is the instantaneous firing rate of presynaptic neuron j, and $\overline{y}_i$ is the trace value of postsynaptic neuron i.

Finally, in order to prevent the unlimited growth of synaptic weights, we renormalized the synaptic weight vector of each postsynaptic neuron i on every numerical timestep as follows

$$\sum_j \left(w_{ij}\right)^2 = 1.$$

2.4. *Preprocessing of camera images for VisNet*

Video frames from the scene camera were extracted using VLC media player. These frames were used as the templates for generating the input images presented to VisNet. The frames were converted to monochrome using the MATLAB function rgb2gray and resized to a 256 x 256 matrix to make it fit to VisNet's retina. We have recently demonstrated how VisNet can be trained using highly realistic images, in which the hand is seen against natural scenes with multiple objects presented simultaneously (Galeazzi et al. 2015). However, this method would require a training regime in which the network is exposed to a large number of different backgrounds. Therefore, since the focus of the current study was on testing the performance of the model incorporating natural movements from human subjects, with the exception of the hand and jigsaw piece, the backgrounds of the frames were filled with a 128 grayscale value as per standard VisNet preprocessing. This procedure also filled missing image regions on VisNet's retina arising as the camera images were shifted in line with eye movements

recorded from the participants. We demonstrated elsewhere (Galeazzi et al. 2015) how the hypothesized learning may operate with the hand presented against complete natural background scenes.

The .csv file generated by the ASL Mobile eye-tracker contained the x and y coordinates of the gaze position of the eyes within the scene frame at a sampling rate of 30 Hz. Missing data values were estimated using linear interpolation. Each of the image frames recorded by the scene camera was shifted across VisNet's retina using these coordinates. In this way, the retinal shifts in VisNet effectively matched the corresponding location of the participant's gaze in the visual scene.

In previous simulations with VisNet it has been shown that the output cells develop single, localized visual receptive fields in localized areas near the hand (Galeazzi et al. 2013, 2015). The network achieves this by extracting the relevant features of the hand and a surrounding object location. In a recent study (Galeazzi et al. 2015) it was shown how the network develops these single, localized visual hand-centered receptive fields after being trained with a variety of objects appearing simultaneously near the hand. However, in the present study we are only exposing the network to images containing a hand and a jigsaw piece. Therefore, in order to prevent VisNet from learning to discriminate between the hand-centered locations by simply exploiting the differences in the shapes of the jigsaw pieces, all experimental trials used pieces with the same shape but with different illustrations. Similarly, to prevent VisNet from differentiating the positions of pieces based on the illustrations, the internal regions of all the pieces were filled with the same grayscale values. A sample of a preprocessed frame is shown in the lower image of Figure 1.

2.5. *Training and testing*

As explained in Section 2.1, on each trial human participants would encounter a jigsaw piece near their hand, and they had to decide where this piece fitted into the puzzle in front of them. There were three possible positions relative to the hand in which the jigsaw piece would appear (i.e. left, right, in front). The participants would typically perform a series of saccades to explore the visual scene while they attempted to solve the puzzle. The images recorded by the scene camera and eye movements recorded by the eye-tracking camera allowed us to reconstruct the retinal images seen by the participant. We then extracted the eye movements before the initiation of the hand movement. These images were processed as described above and then used to train VisNet.

The training procedure for VisNet consisted of presenting the model with successive processed frames extracted from the scene camera. The images would be centered on VisNet's retina according to the location in the scene at

which the participants were fixating. Given the time resolution of the eye-tracker (30 Hz), we updated the retinal position of the image every 33 ms. In order to counterbalance the different training trials within each participant, we adjusted the sequences to the same length. The activations and firing rates of neurons through successive layers of the network were calculated as described above in Section 2.2. The synaptic weights between layers were then updated according to the differential trace rule, see Equation (8).

Five independent VisNet simulations were run, where each simulation used data collected from a different human participant. In each simulation, the processed images from the scene camera were presented shifting on VisNet's retina, where the retinal shifts were guided by the eye movement data collected from the eye-tracking camera. One training epoch for the network consisted of presenting the retinal images constructed from all trials of the participant. In each of the five simulations the network was trained over 50 epochs. Table 1 contains the parameters used for these simulations.

After the network was fully trained on all the trials of the corresponding human participant, the network was tested to determine whether it had developed localized hand-centered receptive fields. Image sequences of the different spatial configurations of the hand and puzzle piece were presented over all the retinal locations where the spatial configuration was previously seen during training. In each simulation, the network was tested with the three spatial configurations of the hand and jigsaw piece corresponding to what the human subjects saw as they tried to solve the puzzle, that is with the jigsaw piece either on the left, or the right, or in front of the hand. Throughout the testing phase, the synaptic weights were unchanged. The neuronal outputs were computed and recorded during the presentation of each spatial configuration in each of the retinal locations.

2.5.1. *Analysis of network performance using information measures*

In order to assess the network's performance, we used single and multiple cell information theoretic measures. This section is reproduced from Galeazzi et al. (2013). These measures can help us determine whether individual cells in the output layer were able to respond to a specific target location in a

Table 1. Parameters used in the computer simulations.

Parameter	Layer 1	Layer 2	Layer 3	Layer 4
Dimensions	256 x 256	256 x 256	256 x 256	256 x 256
#Synapses	200	200	200	200
Learning rate (α)	0.1	0.1	0.1	0.1
Activation time constant (τ_h)	20 ms	20 ms	10 ms	10 ms
Trace time constant ($\tau_{\bar{y}}$)	500 ms	500 ms	500 ms	500 ms
Sigmoid threshold percentile	99.2%	98%	88%	95%
Sigmoid slope (β)	190	40	75	26
Filter radius (σ)	7	11	17	25
Filter contrast (δ)	1.5	1.5	1.6	1.4

hand-centered frame of reference over a number of different retinal locations. In previous VisNet studies, the single cell information measure has been applied to individual cells in the last layer of the network and measures how much information was available from the response of a single cell about which stimulus was shown. Following the conventions of Galeazzi et al. (2013, 2015), in this study a stimulus is defined as one of the three different hand-object configurations (i.e. with the jigsaw piece either on the left, on the right, or in front of the hand). If an output neuron responded to just one of the three spatial configurations, and the cell responded to this configuration across all the tested retinal locations, this meant that the neuron conveyed maximal single cell information. In other words, this cell responded selectively to all the different views of the same hand-centered configuration. The amount of information carried by a single cell about the hand-centered location of an object (jigsaw piece) was computed using the following formula

$$I(s,R) = \sum_{r \in R} P(r|s) \log_2 \frac{P(r|s)}{P(r)} \tag{9}$$

where the stimulus-specific information $I(s,R)$ is the amount of information the set of responses R of a single cell has about a specific stimulus (i.e. object location relative to the hand) s, while the set of responses R corresponds to the firing rate y of a cell to each of the three stimuli (hand-object configurations) presented in all the tested retinal locations. A more detailed description of how the single cell information is calculated is provided in Rolls and Milward (2000), Rolls et al. (1997a) and Rolls (2008).

The maximum single cell information measure is given by

$$\text{Max. single cell info.} = \log_2(\text{Number of stimuli}), \tag{10}$$

where in this case the number of stimuli, that is, spatial configurations of the hand and object, is 3. Thus, this means that the maximum single cell information measure would be 1.58 bits. This is achieved when the cell responds selectively to just one of the three spatial configurations of the hand and object (jigsaw piece), and responds to that spatial configuration over all the tested retinal positions. However, the single cell information value in itself does not provide information regarding which of the three configurations the cell is selective to. An example of how the single cell information is calculated is provided in the Appendix.

The multiple-cell information computed the average amount of information the network has about which hand-centered location the object was presented in obtained from the responses of all the output cells. Using this procedure, we can then verify whether, across the population of output cells,

there was information about all of the three stimuli (i.e. hand-object configurations). From a single presentation of a hand-object configuration, we calculated the average amount of information obtained from the responses of all the cells regarding which hand-object configuration was shown. This was achieved through a decoding procedure that estimates which stimulus s' (i.e, hand-object configuration) gives rise to the particular firing rate response vector on each trial. A probability table of the real stimuli s and the decoded stimuli s' was then constructed. From this probability table, we calculated the mutual information

$$I(S, S') = \sum_{s,s'} P(s, s') \log_2 \frac{P(s, s')}{P(s)P(s')}. \tag{11}$$

For a more detailed description of how the multiple cell information was calculated see Rolls et al. (1997b), Rolls and Milward (2000), and Rolls (2008). Multiple cell information values were calculated for the subset of cells which, according to the single cell analysis, have the most information about which stimulus (i.e., hand-object configuration) was shown. In particular, the multiple cell information was calculated from five cells for each stimulus that had the most single cell information about that stimulus. In the simulations presented in this paper, we showed the network three possible hand-centered object locations, therefore, we performed the multiple cell analysis with a total of 15 cells. In previous research (Stringer and Rolls 2000) it was found that sampling from five cells with maximal single cell information for each stimulus provided a sufficiently large subset of cells to achieve maximal multiple cell information, thus demonstrating that shift invariant representations of each tested hand-object configuration were formed. In other words, each hand-centered location could be uniquely identified.

In addition to the information measures, we provide the firing rates of some of the informative cells in order to provide an example of the type of selective responses observed after training.

3. Results

The fixations were labeled using a semi-automatic fixation detection algorithm (de Urabain et al. 2015). The average fixation duration across all participants was 617 ms. Table 2 shows the average fixation duration for each individual participant as well as the average distance of the fixation points with respect to the hand. We used the chessboard distance metric in pixels to determine the average distance of the fixation points with respect to the hand (Bailey 2004) and subsequently converted this value to degrees of

Table 2. Average fixation durations and average distance of fixations with respect to the hand.

Participants	Fixation duration mean (SD)	Distance from hand mean (SD)
1	534ms (327)	6.412°(3.78)
2	319ms (293)	10.511°(4.57)
3	681ms (592)	6.868°(4.41)
4	728ms (620)	7.784°(3.79)
5	824ms (637)	7.766°(3.76)

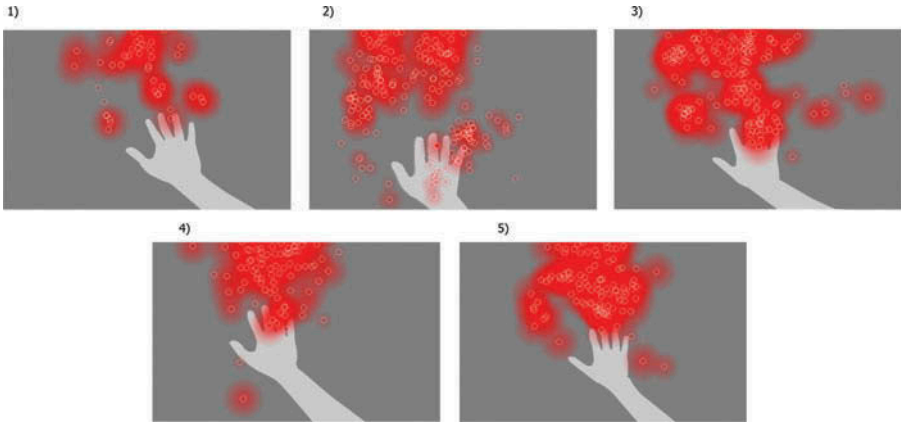

Figure 3. Heat maps showing the fixations made by each subject during the experiment.

visual angle. Fixation heat-maps for each of the participants are shown in Figure 3.

We ran five separate simulations using data recorded from each of the five different human test subjects. In each simulation we analyzed the firing rates of the output layer in VisNet before and after training the network with the footage generated from the scene camera during the jigsaw puzzle task with the human participant. The network was tested before and after training to determine whether cells in the output layer learned to respond selectively to localized hand-centered receptive fields and responded invariantly as these hand-object configurations shifted across the different views.

Figure 4 shows the firing rate responses of three output neurons before training in one of the simulations. Each of the three columns shows the firing rate responses of a particular output cell, which is labeled at the top of the column. The three rows of plots show the responses of the cells to each of the stimuli presented during testing. The top row shows the firing rate responses of the output cells when a jigsaw piece is on the right side of the hand, the row in the middle shows the firing rate responses when the jigsaw piece is on the left side of the hand, and the bottom row corresponds to the jigsaw piece presented in front of the hand. Each individual subplot presents the firing rate responses of the given cell as the particular hand-object configuration is

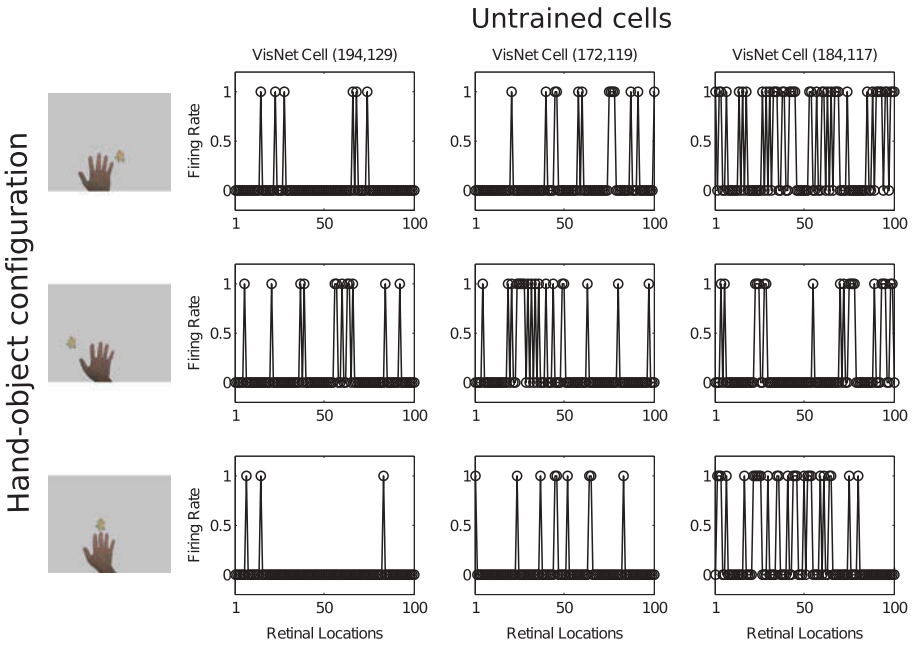

Figure 4. Firing rate responses of three untrained output cells in the top layer of VisNet. The columns shows the firing responses for each of the three output cells. Each row shows the responses of the cells to one of the three hand-object configurations (i.e. with the jigsaw piece either on the left, or the right, or in front of the hand) over 100 tested retinal locations shown along the abscissae. In this Figure we can observe that the cells have unstructured responses across the different views of each of the three hand-centered configurations.

shifted over 100 retinal locations. We can see that before training the three cells responded in an unstructured manner.

Figure 5 shows the responses of the same output neurons after training. Here we can see that after training, each of the output neurons responded much more selectively to one of the hand-object configurations. For example, cell 194, 129 (first column) is perfectly selective in that it responds to the same hand-object configuration across all of the tested retinal locations, and never responds to either of the other two hand-object configurations. The next two columns also show highly selective cells, each of which responds predominantly to one particular hand-object configuration across most or all retinal locations, and responds only rarely to any of the other configurations. The results from the other four simulations were similar.

Additionally, as a global measure of performance, we conducted information analysis on the responses of the output cells to all of the test image sequences. Information analysis was conducted separately for each of five VisNet simulations that were performed using data recorded from the five different human test subjects. For each simulation, we analyzed the responses of the untrained network as well as the responses after the network was trained on the images of hand-object configurations generated from the

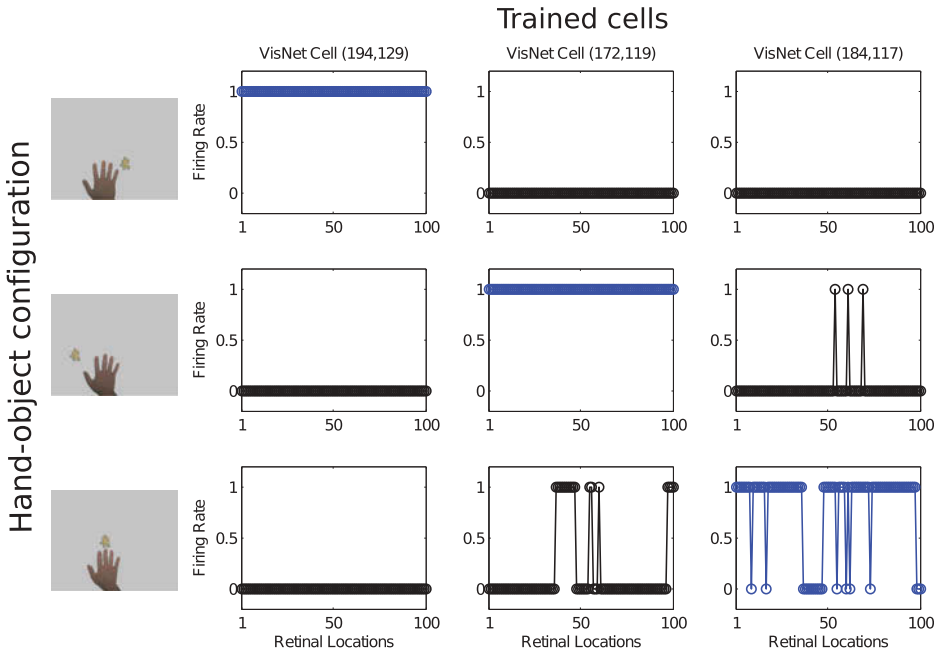

Figure 5. Firing rate responses of the same three output cells shown in Figure. 4 after they have been trained on retinal images of the hand and jigsaw piece recorded as a human subject tries to complete a puzzle. Conventions as in Figure. 4. In this Figure we can observe that the cells show a preference for a particular hand-centered configuration and respond to that configuration over most of the 100 tested retinal positions shown along the x axis of each subplot.

human trials. Figure 6 shows the single and multiple cell information measures for each of the simulations.

In order to determine the robustness of the performance of the trace learning rule, we conducted further simulations in which the trace parameter value ($\tau_{\bar{y}}$) was varied within the range of 100 and 500 ms. In these additional simulations, we continued to obtain robust performance with the model when driven using the data recorded from all five participants. Furthermore, since our experimental paradigm involved collecting data from participants in unrestrained conditions, our results suggest that the learning mechanism is robust over a range of different movement sequences carried out by different subjects.

The single cell information analysis (Figure 6 top) shows that before training the output cells conveyed very little information about the three different hand-object configurations (i.e. with the jigsaw piece either on the left, on the right, or in front of the hand). This means that, as was shown in the response profiles of Figure 4, the cells respond randomly or in a unstructured way to the different hand-object configurations. After training, we found that in all simulations there is a substantial increase in the amount of single cell information. In Figure 6, a subset of cells in each simulation

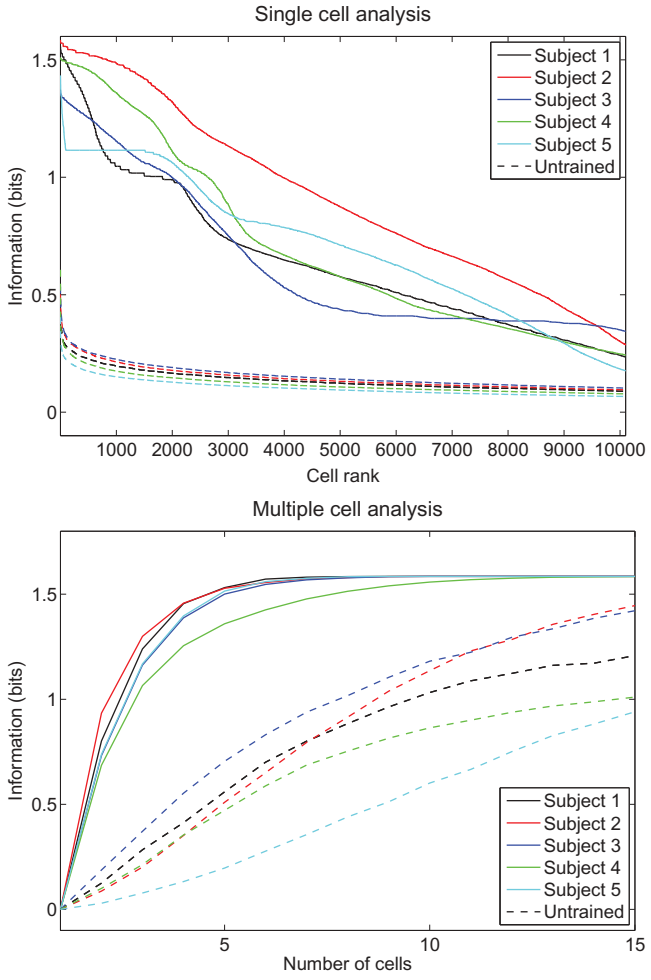

Figure 6. Information analysis for each of five VisNet simulations that were performed using data recorded from each of the five different human test subjects. The five different simulations are represented by different colored plots, where the dashed line shows performance before training and the unbroken line is after training. The upper plot shows the single cell information analysis. This shows that after training, the single cell information conveyed by individual neurons increased substantially in all test cases. Before training, no cells reached the maximal information value of 1.58 bits, whereas in the trained condition a portion of cells in each simulation reached high or even maximal information. These trained cells were highly selective to one of the three tested hand-object configurations, and responded to that configuration in most or all of the tested retinal locations. The lower plot shows the multiple cell information analysis. Here it is shown that in all simulations the maximal information value is reached, confirming that all of the configurations are represented selectively by the output neurons.

reached relatively high levels of information. This suggests that the trained cells developed a high selectivity to a particular configuration of the hand and a puzzle piece, and responded to that hand-centered configuration in most or all of the tested retinal locations.

The multiple cell information analysis (Figure 6 bottom) shows that before training the multiple cell information does not reach the theoretical maximum of 1.58 bits. However, after training we can see that multiple cell information asymptotes to the maximal value in all simulations. This suggests that all three of the hand-object configurations are successfully represented by separate cells in the output layer. In particular, even when individual cells did not encode the maximal amount of single cell information, the multiple cell analysis confirmed that all three hand-object configurations could still be differentiated using just 15 cells that individually carried high levels of single cell information about one of the three configurations.

4. Discussion

The simulations presented in this paper investigated a key issue related to the biological plausibility of the trace learning hypothesis presented in Galeazzi et al. (2013, 2015). The trace learning mechanism is able to bind retinal input images that tend to appear in close temporal proximity. This is achieved by incorporating a memory trace of recent neuronal activity into the synaptic learning rule. Our previous modeling studies assumed that for much of the time humans are continually changing their gaze direction as they explore visual scenes containing fixed spatial configurations of the hand with other objects present Galeazzi et al. (2013, 2015). However, in our earlier modeling studies we simply imposed artificially simulated gaze changes and we never tested the plausibility of our hypothesis driving the network with realistic gaze changes recorded from human test subjects.

Therefore, the goal of the current paper was to explore the validity of our assumption and test whether trace learning could exploit natural eye and head movements of human participants exploring a visual scene in order to produce hand-centered visual representations.

It was found that the reconstructed retinal images, which arose from natural movements of the eyes and head as humans undertook the jigsaw task, were able to drive the development of neurons with hand-centered receptive fields in the output layer of the network model. In our simulations it was found that after training the network, there was an increase in the amount of single cell and multiple cell information that the output cells carried about the different hand-centered locations. The output cells learned to fire selectively to the hand and a puzzle piece in a specific hand-object configuration over a large number of retinal locations. In fact, the multiple cell analysis showed that in all simulations a sample of 15 cells was sufficient to differentiate between the three different hand-object spatial configurations. Thus, this work confirms that our neural network model still develops

hand-centered visual representations when trained on the retinal images of human participants under ecological test conditions.

The relation between the activation time constant and trace time constant is important to enable the trace learning mechanism to work efficiently. In our model, the cells within each layer should have a fast enough time constant (τ_h) in order to allow them to respond in relatively short timescales to the changing retinal images as the participants continually shift their gaze around the visual scene. On the other hand, the trace time constant $\tau_{\overline{y}}$ should be long enough to allow the network to associate together different retinal images corresponding to different retinal locations. Trace learning binds together different input patterns across time, and this requires that the afferent synapses on postsynaptic neurons remain associatively modifiable through a slowly decaying postsynaptic memory trace of recent activity. Our simulations confirmed that the performance of the model was robust for values of $\tau_{\overline{y}}$ across the interval $[100\text{ms}, 500\text{ms}]$. This was found to be the case when the model was driven by data recorded from any of the five participants in unrestrained conditions. Trace learning has been implemented previously in a spiking neural network, where it was shown that the effect of the trace is sensitive to the shortening or lengthening of the activation time constants (Evans and Stringer 2012). Given our understanding of how trace learning operates, we anticipate that the trace learning mechanism would not be impaired by variations in fixation patterns and scan paths, given that the memory trace of the learning rule would still bind together the concurrent input patterns irrespective of the dynamics of visual exploration.

In the simulations reported in this paper, the model was driven by experimentally recorded gaze changes from five human participants over relatively brief time intervals of the order of tens of seconds. This meant that there was uneven coverage of the retinal space. This impairs the ability of the model to respond to every possible hand-centered location invariantly across all retinal locations. More extensive visual training would be necessary in order to establish complete visual coverage of the hand-centered space. However, under more ecological conditions, humans are clearly exposed to much longer timescales of visual training.

This is the first time that the flow of visual training images presented to VisNet is guided by actual gaze changes recorded from human participants, instead of imposing idealized image sequences based on theoretical assumptions about the statistics of eye and head movements. Therefore, these findings are critical to show that the network is able to self-organize hand-centered visual representations successfully with retinal images arising from realistic movements of the eyes, head and hand. Our results suggest that despite the variability of the visual search strategies and gaze changes of individual participants, trace learning was robust enough to allow the network to form neurons with hand-centered receptive fields that are invariant to shifts in retinal

location, as observed in neurophysiology studies (Bremner and Andersen 2012, Buneo and Andersen 2006, Graziano et al. 1997, Graziano and Gross 1998).

Acknowledgments

JMG is supported by The Oxford Foundation for Theoretical Neuroscience and Artificial Intelligence and the Consejo Nacional de Ciencia y Tecnología (CONACYT). JN is supported by the European Research Council StG (NEUROCODEC, #309865). We would also like to thank the Latin American School on Computational Neuroscience (LASCON) for facilitating this collaboration.

References

Andersen RA, Essick GK, Siegel RM. 1985. Encoding of spatial location by posterior parietal neurons. Science. 230(4724):456–458.

Bailey DG. 2004. An efficient Euclidean distance transform. In: Klette R Žunić J, editors. Combinatorial image analysis. Berlin: Springer; p. 394–408.

Bremner L, Andersen R. 2012. Coding of the reach vector in parietal area 5d. Neuron. 75(2):342–351.

Brotchie PR, Andersen RA, Snyder LH and Goodman SJ. 1995. Head position signals used by parietal neurons to encode locations of visual stimuli. Nature. 375(6528):232–235.

Brozzoli C, Gentile G, Ehrsson HH. 2012. That's near my hand! Parietal and premotor coding of hand-centered space contributes to localization and self-attribution of the hand. J Neurosci. 32(42):14573–14582.

Brozzoli C, et al. 2011. FMRI adaptation reveals a cortical mechanism for the coding of space near the hand. J Neurosci. 31(24):9023–9031.

Buneo C, Andersen R. 2006. The posterior parietal cortex: sensorimotor interface for the planning and online control of visually guided movements. Neuropsychologia. 44(13):2594–2606.

Buneo C, Jarvis MR, Batista AP, Andersen RA. 2002. Direct visuomotor transformations for reaching. Nature. 416(6881):632–636.

de Urabain IRS, Johnson MH, Smith TJ. 2015. GraFIX: a semiautomatic approach for parsing low-and high-quality eye-tracking data. Behav Res Methods. 47(1):53–72.

Evans BD, Stringer SM. 2012. Transformation-invariant visual representations in self-organizing spiking neural networks. Front Comput Neurosci. 6:46.

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

Galeazzi JM, Mender BM, Paredes M, Tromans JM, Evans BD, Minini L, Stringer SM. 2013. A self-organizing model of the visual development of hand-centered representations. PloS One. 8(6):e66272.

Galeazzi JM, Minini L, Stringer SM. 2015. The development of hand-centered visual representations in the primate brain: a computer modeling study using natural visual scenes. Front Comput Neurosci. 9:147.

Gentile G, Petkova VI, Ehrsson HH. 2011. Integration of visual and tactile signals from the hand in the human brain: an FMRI study. J Neurophysiol. 105(2):910–922.

Graziano MS, Gross CG. 1998. Spatial maps for the control of movement. Curr Opin Neurobiol. 8(2):195–201.

Graziano MS, Yap GS, Gross CG. 1994. Coding of visual space by premotor neurons. Science. 266(5187):1054–1054.

Graziano M, Hu X, Gross C. 1997. Visuospatial properties of ventral premotor cortex. J Neurophysiol. 77(5):2268–2292.
Makin TR, Holmes NP, Brozzoli C, Rossetti Y, Farne A. 2009. Coding of visual space during motor preparation: approaching objects rapidly modulate corticospinal excitability in hand-centered coordinates. J Neurosci. 29(38):11841–11851.
Makin TR, Holmes NP, Zohary E. 2007. Is that near my hand? Multisensory representation of peripersonal space in human intraparietal sulcus. J Neurosci. 27(4):731–740.
Mender BM, Stringer SM. 2014. Self-organization of head-centered visual responses under ecological training conditions. Netw Comput Neural Syst. 25(3):116–136.
Pesaran B, Nelson M, Andersen R. 2006. Dorsal premotor neurons encode the relative position of the hand, eye, and goal during reach planning. Neuron. 51(1):125.
Rolls ET, Milward T. 2000. A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition and information-based performance measures. Neural Comput. 12:2547–2572.
Rolls ET, Stringer SM. 2007. Invariant global motion recognition in the dorsal visual system: a unifying theory. Neural Comput. 19(1):139–169.
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. 114:177–185.
Rolls ET. 2008. Memory, attention, and decision-making: a unifying computational neuroscience approach. Vol. 1. Oxford: OUP.
Rolls ET, Treves A. 2011. The neuronal encoding of information in the brain. Prog Neurobiol. 95(3):448–490.
Rolls ET, Treves A, Tovee MJ, Panzeri S 1997b. Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. 4:309–333.
Rolls ET, Webb TJ. 2014. Finding and recognizing objects in natural scenes: complementary computations in the dorsal and ventral visual systems. Front Comput Neurosci. 8:85.
Stringer SM, Rolls ET. 2000. Position invariant recognition in the visual system with cluttered environments. Neural Netw. 13(3):305–315.
Stringer S, Perry G, Rolls ET, Proske JH. 2006. Learning invariant object recognition in the visual system with continuous transformations. Biol Cybernet. 94(2):128–142.
Wallis GM, Rolls ET. 1997. Invariant face and object recognition in the visual system. Progr Neurobiol. 51(2):167–194.

Appendix A. Computing single cell information

The single cell information measure used in these simulations is given by Equation (5). Here we provide a numerical example of how these values are computed.

In this case, we will consider single cell information measures for simulations with four different hand-object configurations, A, B, C, and D, and 100 different retinal locations. As each hand-object configuration is presented an equal number of times the probability of each spatial configuration being presented, $P(s)$ will be $P(s) = 1/4$. To calculate the probability of each response the firing rates for each cell, r, are binned. The binning procedures are described in more detail by [and Treves2011]. In VisNet we generally use equispaced bins to apply the information analysis, for example, by using three equally spaced bins, $0 \leq r < 0.33$, $0.33 \leq r < 0.67$, and $0.67 \leq r \leq 1$. We produce a matrix of responses for each cell, an example is given in Table A1.

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

Table A1. Example cell firing rates of an individual cell to each hand-object configuration presented in 100 different spatial locations.

Hand-object configurations	$0 \leq r < 0.33$	$0.33 \leq r < 0.67$	$0.67 \leq r \leq 1$
A	3	17	80
B	68	31	1
C	73	25	2
D	65	12	17

$P(r)$ and the probability of the responses given the stimulus $P(r|s)$. For example, the strongest category of response $0.67 \leq r \leq 1$ has the probability of occurring $P(r) = 100/400 = 0.25$ and the probability of occurring given that configuration A was presented $P(r|s) = 80/100 = 0.8$. Therefore, given Equation (5) the amount of information about configuration 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. If all the responses to a single hand-object configuration fall within the maximal response bin, while all other responses to other hand-object fall in a different bin, then the cell performance is optimal. That is, the stimulus-specific information or surprise, $I(s, R)$, conveyed by this cell is maximal. The maximum single cell information value is given by Equation (10), which denotes the maximum information capacity of a single cell. This maximum value of stimulus-specific information depends on the number of stimuli, which in this case corresponds to the number of possible hand-centered configurations.