Learning View Invariant Recognition with
Partially Occluded Objects

July 28, 2009

**Keywords:** object recognition, continuous transformation, occlusion,
inferior temporal cortex

Acknowledgements: This research was supported by the Wellcome Trust, and
by the Economic and Social Research Council. JMT is an ESRC-supported
graduate student.
Abstract

This paper investigates how a neural network model of the ventral visual pathway, VisNet, can form separate view invariant representations of a number of objects seen rotating together. In particular, in the current work one of the rotating objects is always partially occluded by the other objects present during training. A key challenge for the model is to link together the separate partial views of the occluded object into a single view invariant representation of that object. We show how this can be achieved by Continuous Transformation (CT) learning, which relies on spatial similarity between successive views of each object. After training, the network had developed cells in the output layer which had learned to respond invariantly to particular objects over most or all views, with each cell responding to only one object. All objects, including the partially occluded object, were individually represented by a unique subset of output cells.
Introduction

It is important to understand how invariant representations of individual objects are built in the primate visual system even when multiple objects are present in natural scenes. Neurophysiological research has provided substantial evidence showing that over successive stages, the visual system develops neurons that respond with view, size and position (translation) invariance to objects or faces (Desimone, 1991; Perrett and Oram, 1993; Rolls, 1992, 2000; Rolls and Deco, 2002; Tanaka, Saito, Fukada, and Moriya, 1991). For example, it has been shown that the inferior temporal visual cortex has neurons that respond to faces and objects with translation (Ito, Tamura, Fujita, and Tanaka, 1998; Kobatake and Tanaka, 1994; Op De Beeck and Vogels, 2000; Tovee, Rolls, and Azzopardi, 1994), and view (Booth and Rolls, 1998; Hasselmo, Rolls, Baylis, and Nalwa, 1989) invariance.

Stringer and Rolls (2000) showed that a hierarchical neural network model of the ventral visual pathway, VisNet, could recognise objects presented against natural cluttered scenes, providing the model had been previously trained with each object presented individually transforming against a blank background. However, the network failed to learn to recognise individual objects if the objects were presented against a natural cluttered background during training.

Recent studies by Stringer and Rolls (2008) and Stringer et al. (2007) have shown how VisNet may cope with complex scenes during training, and learn invariant representations of individual objects even when no single object is seen in isolation. These modelling studies used the statistics of the natural environment where features within an object occur together more frequently than features between different objects. Specifically, VisNet could learn invariant representations of individual objects if different combinations of transforming objects were seen at different times.

However, a further challenge is to explain how invariant representations can be learned when the objects are partially occluded by one another during learning. Stringer et al. (2007) proposed that Continuous Transformation
(CT) learning (Stringer et al., 2006) combined with the statistical independence of objects presented in different combinations might allow the network to solve this problem. Specifically, consider presenting a number of objects to the network in different subset combinations, but where one of the objects is always partially occluded by whichever objects it is currently shown with. The hypothesis is that the network will simultaneously form separate representations of all of the different objects, where an invariant representation of the partially occluded object is formed by linking together the different partial views through CT learning. However, Stringer et al. (2007) provided no simulation evidence that this could work. In this paper we demonstrate for the first time this process operating with simulated 3-dimensional rotating objects. It is important to investigate this issue because objects in the natural environment will often overlap. This task is more difficult than simply forming separate representations of different objects because, in order for the network to build a complete invariant representation of the partially occluded object, the network has to link together the different partial views of the object as well as separate these partial views from the other objects present.

In the simulations described below, we show how VisNet can form separate view invariant representations of individual objects seen rotating together, where one of the rotating objects is always partially occluded by the other objects present during training. The network develops cells in the output layer which have learned to respond invariantly to particular objects over most or all views, with each cell responding to only one object. All objects, including the partially occluded object, are individually represented in this way by a unique subset of output cells. This learning process relies on the statistical independence of the objects that are shown in different combinations, as well as an invariance learning mechanism known as Continuous Transformation (CT) learning that is described next.
Method

Continuous Transformation Learning

A leading computational theory of how the ventral visual pathway in the brain may develop neurons that respond to objects with transform (e.g. view or location) invariance is Continuous Transformation (CT) learning. CT learning uses an associative (Hebbian) synaptic modification rule (Stringer, Perry, Rolls, and Proske, 2006) that can exploit the image similarity across successive transforms (e.g. views) of a continuously transforming object in order to develop output neurons which respond to the object over all transforms.

The CT learning process is illustrated in Fig. 1 and operates as follows. The network shown has an input layer where stimuli are presented, and an output layer where transform invariant representations develop through learning. The output layer operates as a competitive network, where individual cells send inhibitory projections to the other cells in this layer, and thereby compete with each other. Initially, the weights of the feedforward synaptic connections are set to random values. Then, during learning, a stimulus is initially presented in position 1 (shown in Fig. 1a) and is represented by three active neurons in the input layer (neurons 1, 2, and 3). Activity propagates through the random feedforward connections to the output layer, where one of the neurons, say neuron 8, wins the competition. The simultaneous activation of neurons in the input and output layers causes the synaptic connections between them to become strengthened according to a Hebbian learning rule

\[ \delta w_{ij} = \alpha y_i x_j \]  

where \( \delta w_{ij} \) is the increment in the synaptic weight \( w_{ij} \), \( y_i \) is the firing rate of the post-synaptic neuron \( i \), \( x_j \) is the firing rate of the pre-synaptic neuron \( j \), and \( \alpha \) is the learning rate. To restrict and limit the growth of each neuron’s synaptic weight vector, \( w_i \) for the \( i \)th neuron, its length is normalised at the end of each timestep during training as is usual in competitive learning (Hertz, Krogh, and Palmer, 1991).
Figure 1: An illustration of how CT learning functions in a feed-forward one-layer network. Activation of overlapping neurons during the transformation of the object from position to position leads to the activation of the same neuron in the output layer. Connections are strengthened according to a Hebbian learning rule after each presentation of the stimulus.
As the stimulus moves from position 1 to position 2 (shown in Fig. 1b), it causes activation in the input layer to also move along one neuron at a time. Therefore, when the stimulus is in position 2, it causes neurons 2, 3 and 4 to become active. The overlap in the input space allows two neurons in the input layer to remain active (neurons 2 and 3) during both transformations. The activation of the same neurons in the input layer causes the same neuron in the output layer (neuron 8) to become active again because the connections have already been strengthened when the stimulus was in position 1. The simultaneous activation of the output neuron, with input neurons 2, 3 and the additional input neuron 4 causes their synaptic connections to become strengthened according to the Hebbian leaning rule. Therefore, the activation of neuron 8 will now become associated with the activation of neurons 2, 3 and 4. As the stimulus continues to move from one position to the next, the process repeats itself and the same neuron in the output layer remains activated. This output neuron becomes a position invariant neuron. A more comprehensive description of Continuous Transformation learning and simulation results in the context of invariant object recognition is provided by Stringer et al. (2006) and Perry, Rolls, and Stringer (2006).

**Objects**

Fig. 2 shows the objects used to train the network. There were $N = 6$ continuously rotating 3D objects on a grey background. Previous research (Stringer and Rolls, 2008) has shown that $N = 6$ objects is sufficient to allow VisNet to develop representations of individual objects when the network was trained on object-pairs. The objects were designed and created using the 3D modelling tool Swift 3D 5.4. Ambient lighting with a diffuse light source was added to allow different surfaces to be shown with different intensities. Each object had a rotation offset of 45° along the x, y and z axis and rotated in 1° steps over 360° around the y-axis. This step size was chosen because past research (Stringer et al., 2006) has revealed that it was sufficiently small for CT learn-
Figure 2: The six objects used to train the network. The objects are 3D objects each shown from 360 different views. The effect of the ambient lighting and single diffuse light source is illustrated. This allows different surfaces to be shown with different intensities. Objects are split into two groups; occluding objects are presented in the top row and the occluded object is presented in the bottom row.

ing to operate. The 360 views of each object were then exported as 2D JPG images and encapsulated as Adobe Shock Wave Files. The objects were then aligned and organised using Adobe Flash CS4.

The stimulus set was comprised of five occluding objects and one occluded object. During each training sequence, the occluded object was shown rotating with one of the occluding objects. The spatial arrangement of the objects is shown in Fig. 3. The occluding objects were presented in a pentagon formation. The occluded object, the Jaimoid (irregular multifaceted three dimensional object, Fig. 4), was always presented in the centre of the pentagon.

Each of the occluding objects was placed at one of the five points of the pentagon, partially overlapping the Jaimoid at the centre. The occluding objects were equidistant from the centre of the occluded object, therefore occluding it to the same extent. The occluded object was always behind the occluding objects and in the middle of the pentagon formation. This spatial formation was chosen because it was necessary to ensure that different parts of the occluded object were covered by the five occluding objects.
Figure 3: The pentagon formation specifying the location of the occluding and occluded objects. The occluded object, the Jaimoid, is always presented in the centre of the pentagon. Each of the occluding objects is placed at one of the five points of the pentagon, partially overlapping the Jaimoid at the centre. This ensures that the occluding objects are equidistant from the centre of the occluded object, therefore occluding it to the same extent.

Figure 4: Five example frames selected from the 360 frame testing image sequence of the Jaimoid rotating through 360° in 1° steps. The selected frames shown are for 0°, 72°, 144°, 218° and 288°. All 5 of the occluding objects were also presented to the network in the same manner.

**The VisNet Model**

The model architecture (VisNet) implemented by Wallis and Rolls (1997) that is used to investigate the properties of CT learning in this paper is based on the following: (i) A series of hierarchical competitive networks with local graded inhibition. (ii) Convergent connections to each neuron from a topologically corresponding region of the preceding layer, leading to an increase in the receptive field size of neurons through the visual processing areas. (iii) Synaptic plasticity based on a Hebb-like learning rule. Model simulations which incorporated these hypotheses with a modified associative learning rule.
Figure 5: Left: Stylised image of the four layer network. Convergence through the network is designed to provide fourth layer neurons with information from across the entire input retina. Right: Convergence in the visual system V1: visual cortex area V1; TEO: posterior inferior temporal cortex; TE: inferior temporal cortex (IT).

to incorporate a short term memory trace of previous neuronal activity were shown to be capable of producing object-selective but translation and view invariant representations (Wallis and Rolls, 1997; Rolls and Milward, 2000; Rolls and Stringer, 2001).

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

The model consists of a hierarchical series of four layers of competitive networks, corresponding to V2, V4, the posterior inferior temporal cortex, and the anterior inferior temporal cortex, as shown in Fig. 5. The forward connections to individual cells are derived from a topologically corresponding region of the preceding layer, using a Gaussian distribution of connection.
probabilities. These distributions are defined by a radius which will contain approximately 67% of the connections from the preceding layer. The values used are given in Table 1.

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

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

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

\[
\Gamma_{xy}(\rho, \theta, f) = \rho \left[ e^{-\left(\frac{x \cos \theta + y \sin \theta}{\sqrt{2}/f}\right)^2} - \frac{1}{1.6} e^{-\left(\frac{x \cos \theta - y \sin \theta}{1.6\sqrt{2}/f}\right)^2} \right] e^{-\left(\frac{x \sin \theta - y \cos \theta}{3.5\sqrt{2}/f}\right)^2}
\]  

(2)

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

| Frequency | 0.5  | 0.25 | 0.125 | 0.0625 |
|-----------|------|------|-------|--------|
| Number of Connections | 201  | 50   | 13    | 8      |

Table 2: Layer 1 connectivity. The numbers of connections from each spatial frequency set of filters are shown. The spatial frequency is in cycles per pixel.

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

$$h_i = \sum_j w_{ij}y_j \quad (3)$$

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

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

$$I_{a,b} = \begin{cases} 
-\delta e^{-\frac{a^2 + b^2}{\sigma^2}} & \text{if } a \neq 0 \text{ or } b \neq 0, \\
1 - \sum_{a\neq 0 \text{ or } b\neq 0} I_{a,b} & \text{if } a = 0 \text{ and } b = 0. 
\end{cases} \quad (4)$$

The lateral inhibition parameters are given in Table 3.

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

Table 3: Lateral inhibition parameters

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

$$y = f_{\text{sigmoid}}(r) = \frac{1}{1 + e^{-2\beta(r-\alpha)}} \quad (5)$$

where $r$ is the activation (or firing rate) after lateral inhibition, $y$ is the firing rate after contrast enhancement, and $\alpha$ and $\beta$ are the sigmoid threshold and slope respectively. The parameters $\alpha$ and $\beta$ are constant within each layer, although $\alpha$ is adjusted to control the sparseness of the firing rates. For example, to set the sparseness to, say, 5%, the threshold is set to the value of the 95th percentile point of the activations within the layer. The parameters for the sigmoid activation function are shown in Table 4.
Training Procedure

The occluded object, the Jaimoid, paired with each of the five surrounding occluding objects, is presented to VisNet with both objects rotating over 360° (Fig. 6). Each full revolution over 360° of the pair is followed by the occluded object paired with a different occluding object in a different location around the pentagon formation. This process is repeated until the occluded object is paired with all five occluding objects (Fig. 7).

In addition, all possible pairings of the five occluding objects are then presented in a similar fashion rotating over 360°. This helped VisNet to learn separate representations of the objects by using the statistics of the natural environment where the features within an object occur together more frequently than features between different objects (Stringer and Rolls, 2008). It should be noted that adjacent pairs of occluding objects would also sometimes overlap during training, leading to one occluding object being partially occluded by another occluding object.

At each presentation, the activation of individual neurons is calculated, then their firing rates are calculated, and the synaptic weights are updated. The occluded object paired with all five occluding objects across all 360 transforms followed by all the possible pairings of the occluding objects rotating over all 360° are presented within one epoch of training.

In this manner, the network is trained one layer at a time starting with layer 1 and finishing with layer 4. Fifty training epochs were used for layers 1-4. The learning rate for layers 1-4 were 0.109, 0.1, 0.1 and 0.1, respectively. The population sparseness of the output firing rates was set to 0.992, 0.98, 0.81, 0.91 for layers 1 to 4, respectively.
Figure 6: Five example frames selected from the 360 frame training image sequence of the Jaimoid and the Cone rotating through 360° in 1° steps. The selected frames shown are for 0°, 72°, 144°, 218° and 288°. The Jaimoid is also occluded by the four other occluding objects in separate image sequences. Therefore, in total, there are 5 image sequences used during training, each containing 360 frames.

Figure 7: Five example frames of the Jaimoid occluded by all five occluding objects in their five corresponding positions. The occluding objects are arranged around the pentagon formation so that they are equidistant from the centre of the Jaimoid.

**Testing Procedure**

In order to test whether VisNet had built an invariant representation of the central partially occluded object, we presented the occluded object individually, rotating over 360° (Fig. 4). The surrounding five occluding objects were also presented singly in a similar fashion to verify that VisNet had build invariant representations of these objects too.

The networks’ performance to recognise which object is shown during testing is assessed using two information theoretic measures: single and multiple cell information. Full details on the application of these measures to VisNet are given by Stringer et al. (2006). These measures reflect the extent to which cells respond invariantly to an object over a number of different views (trans-
forms), but respond differently to different objects. The single cell information measure is applied to individual cells in layer 4, and measures how much information is available from the response of a single cell about the stimulus that was presented. The single cell information measure for each cell shows the maximum amount of information that the cell conveys about any one object. This is computed using the following formula with details provided by Rolls, Treves, Tovee, and Panzeri (1997) and Rolls and Milward (2000). The object-specific information $I(s, R)$ is the amount of information the set of responses R has about a specific object s, and is given by

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

where $r$ is an individual response from the set of responses $R$. However, the single cell information measure cannot give a complete assessment of VisNet’s performance with respect to invariant object recognition. If the amount of information provided by a single cell is not sufficient to differentiate between which objects are present during testing, the network may have failed to learn, or a distributed representation may have formed that needs information from a population of neurons to encode which object is present. Furthermore, if all output cells learned to respond to the same object then there would in fact be relatively little information available about the set of objects $S$, and single cell information measures alone would not reveal this. To address these issues, we also calculate a multiple cell information measure, which assesses the amount of information that is available about the whole set of objects from a population of neurons.

Procedures for calculating the multiple cell information measure are described by Rolls et al. (1997) and Rolls and Milward (2000). In brief, from a single presentation of an object, we calculate the average amount of information obtained from the responses of all the cells regarding which object is shown. This is achieved through a decoding procedure that estimates which object $s'$ gives rise to the particular firing rate response vector on each trial. A probability table of the real objects $s$ and the decoded objects $s'$ is then
constructed. From this probability table, the mutual information is calculated as
\[
I(S, S') = \sum_{s,s'} P(s, s') \log_2 \frac{P(s, s')}{P(s)P(s')}. \tag{7}
\]

Multiple cell information values are calculated for the subset of cells which, according to the single cell analysis, have the most information about which object is shown. In particular, the multiple cell information is calculated from the first five cells for each object that had the most single cell information about that object. This results in a population of 30 cells given that there were six objects. Previous research (Stringer and Rolls, 2000) found this to be a sufficiently large subset to demonstrate that invariant representations of each object presented during testing were formed, and that each object could be uniquely identified. If the multiple cell information approaches \(\log_2 N\) bits and is large relative to the information available in an untrained network, it means that enough information is present between the members of the subset to discriminate between all of the objects.

**VisNet Simulation Results**

After the network had been trained on pairings of the occluded and five occluding objects, we tested whether the network had built transform invariant representations of the objects through a CT learning effect. By presenting the rotating objects individually (Fig. 4) to the network we were able to record the cell response properties of the neurons in the 4th layer of VisNet for each of the objects used during training.

Populations of cells that responded invariantly to the individual objects were found. These cells responded to only one object and to no views of any of the other objects. Figs. 8 & 9 show the cell response plots for cell (4, 17), selected at random, as each object is rotated through 360° in 1° steps. Fig. 8 shows the responses of the cell before training and Fig. 9 shows the cell responses after training. The six response plots of cell (4, 17) before training
Figure 8: The firing rate responses of cell (4, 17) in the 4th (output) layer of VisNet to the central occluded object (Jamoid) and the five surrounding occluding objects as they rotated through 360° in 1° steps before training. It can be seen that the cell responds randomly to different views of different objects.
Figure 9: The firing rate responses of cell (4, 17) in the 4th (output) layer of VisNet to the central occluded object (Jaimoid) and the five surrounding occluding objects as they rotated through 360° in 1° steps after training. It can be seen that the cell’s response pattern has changed compared to a network before training (Fig. 8). This cell responds to at least 80% of the views of Jaimoid, and to none of the views of the other objects.

show that the cell responds at random to the six objects. After training, the cell has learned to respond to the central occluded object, the Jaimoid, over more than 80% of the view space and does not respond to any view of any of the other objects.

Figs. 10 & 11 show the cell response plots for cell (19, 1) before and after training, respectively. Before training, the cell responds to the objects quite randomly. After training this cell has learnt to respond invariantly to all 360 views of the Dodecahedron, which was one of the occluding objects, and to no views of any other objects. Furthermore, although not shown here, other output cells learned to respond in a selective and invariant manner to each of the other occluding objects. Thus, all of the objects were represented
Figure 10: The firing rate responses of cell (19, 1) in the 4th (output) layer of VisNet to the central occluded object (Jamoid) and the five surrounding occluding objects as they rotated through 360° in 1° steps before training. It can be seen that the cell responds randomly to different views of different objects.
Figure 11: The firing rate responses of cell (19, 1) in the 4th (output) layer of VisNet to the central occluded object (Jaimoid) and the five surrounding occluding objects as they rotated through 360° in 1° steps after training. It can be seen that the cell’s response pattern has changed compared to a network before training (Fig. 10). This cell has become an exclusive invariant cell for the Dodecahedron. It responds invariantly to all 360 views of the Dodecahedron and to no views of any other object.

However, for the occluded object and a few of the occluding objects, while cells learned to respond invariantly to large parts of the view space, no single cells were found to respond to all 360 views. Instead, Fig. 12 shows that a complete 360° representation of the occluded object was encoded by a small population of cells. This figure shows the firing rate response profile of three 4th layer neurons which have learned to respond to the Jaimoid. Plot (a) shows the response profile of cell (4, 17) which has learned to respond to the Jaimoid over most of the views. Plot (b) shows the response profile of cell (32, 2), which has learned to respond to the Jaimoid over an overlapping region of
Figure 12: The firing rate responses of three cells in the 4th (output) layer of VisNet. These cells were found to respond exclusively to the Jaimoid, where they fire to certain views of the Jaimoid but to no views of any other objects. Plot (a) shows the response profile of cell (4, 17) which has learned to respond to the Jaimoid over most of the views. Plot (b) shows the response profile of cell (32, 2), which has learned to respond to the Jaimoid over an overlapping region of the view space including some points that the previous cell did not respond to. Similarly, plot (c) shows the profile of cell (6, 1), which has learned to respond to the Jaimoid over the remainder of the view space that the two previous cells did not respond to.
the view space including some points that the previous cell did not respond to. Similarly, plot (c) shows the profile of cell (6, 1), which has learned to respond to the Jaimoid over the remainder of the view space that the two previous cells did not respond to. Since all three cells presented respond to the Jaimoid exclusively (they don’t respond to any views of any other objects), the firing of either of the three cells would allow a subsequent brain area to recognise that the Jaimoid is being viewed.

Single cell information analysis was run to confirm whether the network had developed cells that responded invariantly to their preferred object (Fig. 13). The unbroken line represents the results obtained after presenting the six objects to a network trained on the 360 views of all possible object pairs, whereas the dashed line represents the results after presenting the six objects to a random untrained network. Single cell information measures for the 4th layer neurons ranked in order of their invariance to the objects are shown. It can be seen that training the network on the object pairs has lead to many of the 4th layer neurons attaining the maximal level of single cell information of 2.58 bits. These neurons have learnt to respond to all of the views of their preferred object. However, it is unclear whether all of the six objects are individually represented by a unique subset of invariant output cells. Indeed, it is possible that these cells are responding to the same object and are therefore unable to provide information regarding which object is present. To ensure that there are cells that respond preferentially to each of the six objects multiple cell information analysis was performed.

Fig. 14 shows the multiple cell information analysis obtained when VisNet was tested with the six individual objects rotating through 360° in 1° steps. Results are presented after training the network (unbroken line), and with random untrained network (dashed line). After the network was trained, 2.35 bits of information was reached (substantially higher than 1.60 bits reached by the untrained network) suggesting that the single cell information results included cells that preferentially responded to all 6 objects and therefore con-
Figure 13: Single cell information results obtained when VisNet was tested with the occluded object and five occluding objects rotating through 360° in 1° steps. Results are presented after training the network (unbroken line), and with random untrained network (dashed line). The single cell information measure for all 4th layer neurons ranked in order of their invariance to the objects is shown. It can be seen that training the network on the object pairs has led to many 4th layer neurons attaining the maximum level of single cell information of 2.58 bits. These cells have learned to respond selectively to individual objects invariantly over all views.
Figure 14: Multiple cell information results obtained when VisNet was tested with the occluded object and five occluding objects rotating through $360^\circ$ in $1^\circ$ steps. Results are presented after training the network (unbroken line), and with random untrained network (dashed line). After the network was trained, 2.35 bits of information was reached, which was substantially higher than 1.60 bits reached by the untrained network. This confirmed that, after training the network, there were cells that responded preferentially to each one of the six objects. This confirmed that the network had created separate view invariant representations of both the occluded and the five occluding objects.

firming that the network has created view invariant representations of both the occluded and the occluding objects. This was also confirmed by inspection of the cell response plots as shown in Figs. 9 and 11.
Discussion

An important question in natural vision is how the brain forms invariant representations of objects that are always partially occluded by other objects during learning. Stringer and Rolls (2008) have shown that a biologically plausible competitive neural network with distributed representations (VisNet) can develop invariant representations of individual objects when no single object is seen in isolation. The network used the statistics of the natural environment where the features within an object occur together more frequently than features between different objects.

In this paper we demonstrate how such a network might form an invariant representation of an object that is always partially occluded by other objects. The mechanism employed for invariance learning is Continuous Transformation (CT) learning. CT learning uses the spatial continuity between the views of individual objects as they transform in the real world, combined with associative learning of feedforward connection weights.

It was found that, after training the network with a rotating object that is always partially occluded, the network is able to form view invariant representation of the object. Therefore the network was able to link together different partial views of the occluded object that were presented during training to create a complete and view invariant representation of the object. It was also found that view invariant representations are also formed for all five occluding objects. This is a challenging task since the occluding objects were always overlapping the occluded object and therefore VisNet had to learn to separate the objects. This is the first time such learning has been shown to happen in a biologically accurate model of the ventral visual system.

In the simulations described above, individual cells learned to respond to only one of the objects. This is known as ‘Grandmother cell’ encoding. However, individual objects were represented by populations of cells that responded invariantly to different parts of the view space and together represented the object over all 360 views. This means that individual objects are
represented by a small unique population of cells over the entire 360° view space. It is also able to separate out the occluding and the occluded objects even though they overlap during training.

Typical non-biologically inspired artificial visual systems often rely on template matching in order to recognise a partially occluded object (Ullmann, 1992; Ying and Castanon, 2002; Do, Lozo, and Jain, 2005). They require a complete template of the object before they can perform successfully in a partially occluded object recognition task. This type of artificial visual systems usually search for specific markers in the partially occluded object that match the template in order to recognise it. Their performance may become fragile if key parts of the template were missing, as would be in the case if these systems were trained on partially occluded objects. What differentiates the current experiment from other experiments in non-biologically inspired artificial vision systems dealing with partially occluded object recognition, is that in the current paper we demonstrated how the biologically inspired model VisNet can create a view invariant representation of an occluded object after being trained only on partial views of it. It was shown that VisNet can perform robustly in an experimental paradigm that typical non-biologically inspired artificial visual systems might find challenging, because it is able to link together the different partial views of the object as well as separate these partial views from the occluding objects. Furthermore, VisNet does not rely on particular features being present during object recognition, because it acts as an associative lookup system.

In this paper we demonstrate that a biologically plausible neural network of the visual system (VisNet) can solve a problem that other artificial vision models find challenging. VisNet does not rely on a template match solution where a small number of key markers on an object are used in order to identify it. Instead, it uses the properties of distributed representation, generalisation and completion (Rolls and Deco, 2002), to recognise the objects. While artificial vision systems might struggle to identify an object when some of its
markers are partially occluded by another object, this does not pose a problem for VisNet, made evident by its ability to form view invariant representations of an object without ever seeing it unobstructed by other objects.
References

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

D. I. Perrett, M. W. Oram, Neurophysiology of shape processing, Image and Vision Computing 11 (1993) 317–333.

E. T. Rolls, Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas, Philos. Trans. R. Soc. Lond., B, Biol. Sci. 335 (1992) 11–20.

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

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

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

M. Ito, H. Tamura, I. Fujita, K. Tanaka, Size and position invariance of neuronal response in monkey inferotemporal cortex, J Neurophysiol 73 (1998) 218–226.

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

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

M. J. Tovee, E. T. Rolls, P. Azzopardi, Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque, J. Neurophysiol. 72 (1994) 1049–1060.
M. C. Booth, E. T. Rolls, View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex, Cereb. Cortex 8 (1998) 510–523.

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

S. M. Stringer, E. T. Rolls, Position invariant recognition in the visual system with cluttered environments, Neural Networks 13 (2000) 305–315.

S. M. Stringer, E. T. Rolls, Learning transform invariant object recognition in the visual system with multiple stimuli present during training, Neural Networks 21 (7) (2008) 888–903, ISSN 0893-6080, doi: http://dx.doi.org/10.1016/j.neunet.2007.11.004.

S. M. Stringer, E. T. Rolls, J. M. Tromans, Invariant object recognition with trace learning and multiple stimuli present during training, Network 18 (2007) 161–187.

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

J. Hertz, A. Krogh, R. G. Palmer, Introduction to the theory of neural computation, Addison Wesley, Workingham, UK, 1991.

G. Perry, E. T. Rolls, S. M. Stringer, Spatial vs temporal continuity in view invariant visual object recognition learning, Vision Res. 46 (2006) 3994–4006.

G. Wallis, E. T. Rolls, Invariant face and object recognition in the visual system, Prog. Neurobiol. 51 (1997) 167–194.

E. T. Rolls, T. Milward, A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and
information-based performance measures, Neural Comput 12 (2000) 2547–2572.

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

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

J. R. Ullmann, Analysis of 2-D Occlusion by Subtracting Out, IEEE Trans. Pattern Anal. Mach. Intell. 14 (4) (1992) 485–489, ISSN 0162-8828, doi: http://dx.doi.org/10.1109/34.126808.

Z. Ying, D. Castanon, Partially Occluded Object Recognition Using Statistical Models, International Journal of Computer Vision 49 (1).

Q. V. Do, P. Lozo, L. C. Jain, A Vision System for Partially Occluded Landmark Recognition, vol. 3809/2005, Springer Berlin / Heidelberg, 2005.