Vision Research 51 (2011) 553–562

Contents lists available at ScienceDirect

Vision Research

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

The role of independent motion in object segmentation in the ventral visual stream: Learning to recognise the separate parts of the body

I.V. Higgins *, S.M. Stringer

Oxford Centre for Theoretical Neuroscience and Artificial Intelligence, Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, England, UK

ARTICLE INFO

Article history:
Received 16 June 2010
Received in revised form 27 January 2011
Available online 12 February 2011

Keywords:
Visual object segmentation
Recognising body parts
Independent motion
Biological motion
Continuous transformation
Inferior temporal cortex

ABSTRACT

This paper investigates how the visual areas of the brain may learn to segment the bodies of humans and other animals into separate parts. A neural network model of the ventral visual pathway, VisNet, was used to study this problem. In particular, the current work investigates whether independent motion of body parts can be sufficient to enable the visual system to learn separate representations of them even when the body parts are never seen in isolation. The network was shown to be able to separate out the independently moving body parts because the independent motion created statistical decoupling between them.

1. Introduction

Visual segmentation of objects into parts is an important perceptual process. It is reported to happen spontaneously and pre-attentively (Biederman, 1987; Cave & Kosslyn, 1993; Hoffman & Richards, 1984; Hoffman & Singh, 1997; Kurbat, 1994; Lamberts & Freeman, 1999; Lamote & Wagemans, 1999; Scholl, 2001; Singh, Seyranian, & Hoffman, 1999; Van Lier & Wagemans, 1998). Neurophysiological research has investigated which brain areas might be responsible for visual object segmentation into parts, and in particular segmentation of living organisms into body parts. For example, single cell recordings in inferior temporal cortex in monkeys revealed cells that responded to separate body parts (Pinsk, DeSimone, Moore, Gross, & Kastner, 2005). Furthermore, similar body part selective areas were identified in the human fusiform gyrus and lateral occipitotemporal cortex (Pinsk et al., 2009; Schwarzlose, Baker, & Kanwisher, 2005) and latest research has identified a separate brain area specific to body parts, the extrastriate body area in the lateral occipitotemporal cortex (Downing, Chan, Peelen, Dodds, & Kanwisher, 2006; Downing, Jiang, Shuman, & Kanwisher, 2001). Therefore the evidence suggests that visual objects are segmented into parts through a spontaneous pre-attentive process, this segmentation is essential to humans and animals, and neurons that respond to parts of objects exist in the brain.

Even though a number of theories have been put forward trying to explain how visual segmentation of objects into parts may operate, it is still not known in detail how representations of parts are formed in the brain. The existing theories of visual segmentation can be divided into two broad groups. The first one suggests that a limited set of predefined primitives exist, and all objects are made of different combinations of them. Therefore objects are made out of these shapes in the same way as words are made out of letters. These shapes are thought to be generalised cylinders (Marr, 1982) or geons (Biederman, 1987). Such theories have however been criticised, since there are examples of segmentation that do not correspond to these proposed intuitive parts and some of the intuitive parts cannot be derived using any existing models of visual object segmentation into parts (De Winter & Wagemans, 2006).

The second group of theories uses geometric rules based on the shape of the objects in order to separate them into parts (Hoffman & Richards, 1984; Siddiqi, Tresness, & Kimia, 1996; Singh et al., 1999). These theories suggest that object parts may emerge through grouping or segmentation processes based on natural constraints of the stimuli. Geometrical theories are based on the principle of singularity that states that a 3D concave crease almost always results in a 2D concave discontinuity. This leads to most studies using 2D outline stimuli. For example, one of the predominant principles of geometrical theories is the principle of transversality that proposes that a sharp concavity on the on the surface of an object is a likely segmentation point between two object parts (Hoffman & Richards, 1984). An example would be the sharp concavity at the shoulder joint where an arm attaches to the body

* Corresponding author.
E-mail addresses: irina.higgins@psy.ox.ac.uk (I.V. Higgins), simon.stringer@psy.ox.ac.uk (S.M. Stringer).
URL: http://www.oftnai.org (I.V. Higgins).

0042-6989/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2011.01.016

trunk (Siddiqi et al., 1996). Such information as depth, colour, shading or texture is ignored. Although the geometrical theories can be quite successful in predicting how participants might segment 2D shapes into constituent parts, none of them can account for all object segmentation.

None of the theories described above based on primitive shapes and geometrical rules answer the 'how' question, as in explain how the process of visual object segmentation into parts happens in the brain. In particular how the interaction between the anatomy and physiology of the visual brain areas with the visual stimuli results in the formation of separate representations of the whole objects as well as their constituent parts. Neurophysiological studies in macaque monkeys (Pasupathy & Connor, 2001, 2002) have found evidence that neurons in area V4 of the ventral visual pathway respond to concave and convex boundary elements, therefore providing evidence for the geometrical theories of object segmentation. Although these studies explain where in the brain different features of objects such as concave and convex boundary elements are represented, they still do not answer the 'how' question defined above. It is still necessary to explain how these cells develop their response properties through synaptic learning driven by neural activity evoked by visual stimuli during the critical period of visual development (Hubel & Wiesel, 1963).

There is research in visual perception that suggests that the ability to segment objects into constituent parts might be acquired through learning and not innate. For example Maurer and Salapatek (1976) found that newborns fail to fixate on stationary embedded contours of internal features in objects (such as eyes on a face), an ability that appears in 2 month old infants. Contour information is important for the geometrical rules of object segmentation into parts, however the research by Maurer and Salapatek (1976) suggests that newborns might not be sensitive to contours innately and that this ability might be acquired through visual experience instead.

Similarly to the problems associated with the geometrical rules theories, theories based on primitive shapes do not explain how these shapes might be acquired, where in the brain they might be stored and how they might affect visual processing either. This paper is set to try and explain how the interaction between the visual input and the anatomy and physiology of the relevant brain areas results in visual object segmentation into parts, therefore providing an insight into the mechanisms underlying the geometrical rules and primitive shapes theories.

Downing et al. (2001), who discovered the body specific part of the visual cortex, the extrastriate body area, have suggested that it might be one of the distinct modules that govern visual processing, similar to the fusiform face area. They suggested that different classes of objects were processed using these different specialised modules with different underlying neural mechanisms. Therefore it was argued that learning to recognise body parts might be different to learning to recognise other objects. Further evidence for this comes from research suggesting that semantic knowledge of the human body might be distinct from the knowledge of any other object categories (Shelton, Fouch, & Caramazza, 1998). Furthermore functional neuroimaging and single-neuron recording studies have found that the superior temporal sulcus is implicated in processing biological motion as well as body representations (Grossman et al., 2000; Howard et al., 1996; Jellema, Baker, Wicker, & Perrett, 2000; Puce, Allison, Bentin, Gore, & McCarthy, 1998; Wachsmuth, Oram, & Perrett, 1994). This might suggest that biological motion might play a role in building representations of separate body parts. Chan, Kravitz, Truong, Arizpe, and Baker (2010) also suggested that representations of body parts in the brain depended on life-long experiences and reflected the statistics with which the stimuli occured.

This paper will use an existing biologically plausible computational model of the ventral visual stream, VisNet, to test a new theory of how visual object segmentation into parts might happen, in particular how separate representations of body parts might develop. This theory is suggested as an explanation to the geometrical rules and the primitive shapes theories, as in how the relevant brain areas learn to perform visual object segmentation into body parts. It is important to understand how representations of separate body parts are formed in the brain, and it is hypothesised that independent motion might enable the brain to do so.

One explanation to how separate representations of body parts can be built in the visual system is the "biased competition hypothesis" of attention. It suggests that in order to build separate representations of individual body parts, feedback connections are necessary, because they provide the mechanism for attentional selection, which isolates individual body parts to enable separate representations of them to be formed (Rolls & Deco, 2002). However, it has been reported that visual object segmentation into parts happens pre-attentively (Biederman, 1987; Cave & Kosslyn, 1993; Hoffman & Richards, 1984; Hoffman & Singh, 1997; Kurbat, 1994; Lamberts & Freeman, 1999; Lamote & Wagemans, 1999; Scholl, 2001; Singh et al., 1999; Van Lier & Wagemans, 1998). Therefore the "biased competition hypothesis" cannot be used to explain how the visual system learns to segment bodies into separate parts.

Research has shown that it is possible to learn about individual objects when multiple objects are present in the scene without the need for an attentional mechanism using purely feedforward connectivity in a hierarchical neural network model of the ventral visual pathway, VisNet (Stringer & Rolls, 2008; Stringer, Rolls, & Tromans, 2007). Stringer and Rolls (2008) showed how the statistical properties of the visual input stimuli play a crucial role in enabling the network to develop view invariant representations of individual objects when multiple objects are present during training. This was achieved through statistical decoupling because features within individual objects occurred more frequently together than features between different objects. Stringer and Rolls (2008) showed that, when training on all possible pairs of objects, at least six objects were necessary in the training set for this statistical decoupling to occur and there was no need for top-down attentional influences. Therefore this study will assume that no attentional influences are exerted on the ventral visual stream to aid visual object segmentation into body parts and no feedback connections will be used in the simulations.

It is expected that similar statistical decoupling to the one described by Stringer and Rolls (2008) might enable separate representations of body parts to develop in the ventral visual stream if the target body parts are engaged in independent motion. However the problem of learning about individual body parts is harder than learning about individual objects in a complex scene. In the study of Stringer and Rolls (2008), individual objects were presented in different combinations with each other, which enabled the statistical decoupling to happen. In the case of body parts no such reconfiguration can happen, as all body parts are always present together. However it is predicted that the effect of independent motion might be similar to the effect of different object couplings in the study of Stringer and Rolls (2008). This is because although the body parts are always seen together, as long as the body parts are moving independently, there will be statistical decoupling between the positions in which the different body parts are seen. This forces the output layer to form separate representations of the different body parts.

In the simulations described below two body parts, two arms will be used. It is expected that independent motion of the two arms will ensure that each transform of one arm will be seen with all the possible transforms of the other arm. Therefore the features

within one transform of an arm will be seen together more often than features between arms. Therefore independent motion is expected to help the network to learn about the individual arms as well as still form a representation of the whole object consisting of the two arms.

It is predicted that when two body parts (in the simulations described below they were arms) are moving together, neurons in the output layer of the network will fail to separate them out. These neurons will respond with transform invariance to both arms. However when the two arms start to move independently, VisNet will be able to separate them out through statistical decoupling. It is expected that there will be two orthogonal populations of cells in the output layer, each of which will respond just to their preferred arm. It is also expected that there will be a population of cells that will still respond to both arms and thus will have learnt to respond to the whole object consisting of the two arms in this instance.

2. Method

2.1. Learning to recognise independently moving body parts

The problem addressed in this paper is how VisNet can form separate representations of two arms always presented together during training. Previous research (Stringer & Rolls, 2008) has shown that in order for the network to learn separate representations of objects when different combinations (e.g. pairs) of multiple objects are present during training, the stimulus set has to contain at least six objects. In this way features within objects will appear together more often than features between objects and statistical decoupling will enable the network to form separate representations of the objects. However, in the present study the training set consists of two arms only and they are always presented together.

In the study of Stringer and Rolls (2008), the objects were rotating together over 90°, therefore no independent motion was present. It is hypothesised that introducing independent motion to the two arms in the present study will act in a similar way to increasing the number of objects in the training set to 6 or more in the study of Stringer and Rolls (2008). Each view of one arm will be presented with each view of the other arm, so that features within a transform of an arm will be presented together more often than features between the two arms. This will force some individual neurons in the output layer to learn to respond to either one arm or the other. Cells which respond to both arms will also remain and they will have learnt to respond to the whole object.

2.2. Objects

The arms were represented by two black rounded rectangles rotating around the cental fixed hinge point in 40 equal steps on a grey background as shown in Fig. 1.

To study whether the lack of independent motion leads to the network failing to learn about the two arms individually, the arms were animated to rotate simultaneously in lock-step starting from the top (Fig. 1). This ensured that each view of one arm was always seen with the same view of the other arm during rotation. This resulted in 40 frames of animation.

Independent motion was animated by creating 1600 frames whereby each of the 40 translations of one arm was presented with each of the 40 translations of the other arm as the arms were rotating in 40 equal steps.

The two sets of stimuli described above were created and animated using Adobe Flash CS4. The frames were exported as JPG images.

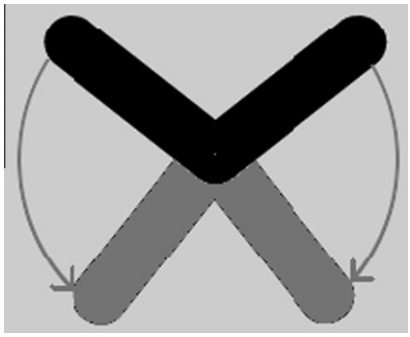

Fig. 1. Two black rounded rectangles representing arms rotating around the cental fixed hinge point in 40 equal steps. The black shapes represent the top locations and the grey shapes represent the bottom locations of the arms. In the first simulation the arms were rotating simultaneously in lock-step so that each rotation of one arm was always seen with the same rotation of the other arm during training. In the second simulation the two arms moved independently, so that all 1600 (40×40) combinations of transforms of the two arms were seen during training.

2.3. Transformation invariance 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, & 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 a large number of transforms. Because CT learning is based on the standard Hebbian learning rule, it is biologically plausible.

An idealised version of the CT learning process outlining the theoretical principle is illustrated in Fig. 2 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. 2a) 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

$$\delta w_{ij} = \alpha y_i x_j \tag{1}$$

where δw_{ij} is the increment in the synaptic weight w_{ij}, y_i is the firing rate of the post-synaptic neuron i, x_j is the firing rate of the pre-synaptic neuron j, and α is the learning rate. To restrict and limit the growth of each neuron's synaptic weight vector, $\mathbf{w}_i$ for the ith neuron, its length is normalised at the end of each timestep during training as is usual in competitive learning (Hertz, Krogh, & Palmer, 1991). This is necessary to ensure that one or a few neurons do not always win the competition. If there was no normalization of synaptic weights during a simple Hebbian learning procedure, just a few neurons may eventually learn to respond strongly to nearly all of the input patterns. Neurophysiological evidence for synaptic weight normalization is provided by Royer and Parè (2003).

As the stimulus moves from position 1 to position 2 (shown in Fig. 2b), it causes activation in the input layer to also move along

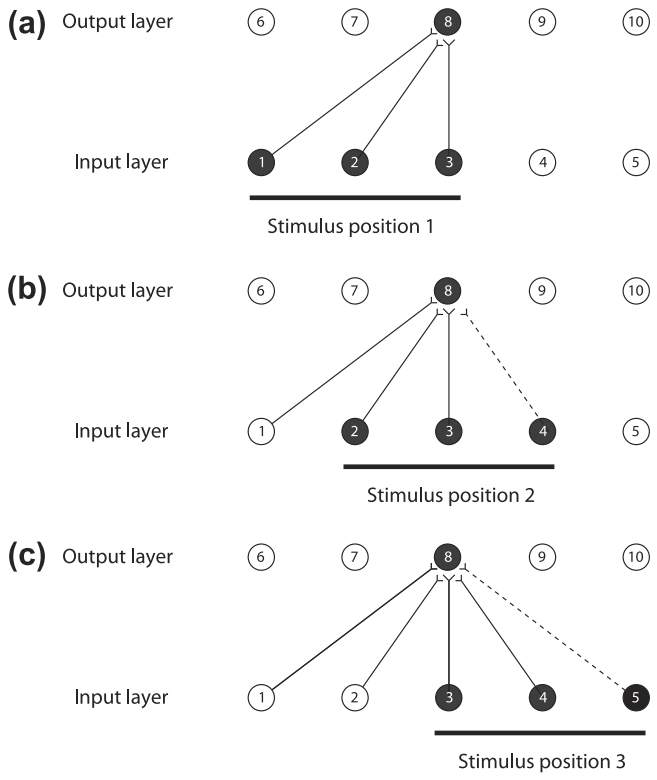

Fig. 2. An illustration of how CT learning functions in a feedforward one-layer network. Activation of overlapping neurons during the transformation of the object from position to position leads to the activation of the same neuron in the output layer. Connections are strengthened according to a Hebbian learning rule after each presentation of the stimulus.

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 et al. (2006).

An alternative to CT learning is trace learning, which has been applied to the problem of visual invariance before by Foldiak (1991) and Wallis et al. (1993). The trace rule is designed to enable neurons to learn from the temporal statistics of the natural visual inputs, which in short time periods are likely to be about the same object. However the simulations described in this paper rely on spatial and not temporal continuity of the transforms of the stimuli and therefore the CT learning rule was chosen.

Another alternative to CT learning is the Slow Features Analysis (SFA) algorithm proposed by Wiskott and Sejnowski (2002) as a mechanism of visual invariance learning. However unlike CT learning, which is based on a standard Hebbian learning rule, the SFA algorithm is not biologically plausible as the authors do not explain how the proposed computations might be performed by neurons in the brain.

2.4. The VisNet model

The VisNet model architecture that is used 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 have shown VisNet to be capable of producing object-selective but translation and view invariant representations (Rolls & Milward, 2000; Rolls & Stringer, 2001; Stringer et al., 2006; Wallis & Rolls, 1997).

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. 3. The forward connections to individual cells are derived from a topologically corresponding region of the preceding layer, using a Gaussian distribution of connection probabilities. These distributions are defined by a radius which will contain approximately 67% of the connections from the preceding layer. The values used are given in Table 1.

Before the objects are presented to the network's input layer they are pre-processed by a set of input filters which accord with the general tuning profiles of simple cells in V1. The filters provide a unique pattern of filter outputs for each transform of each visual

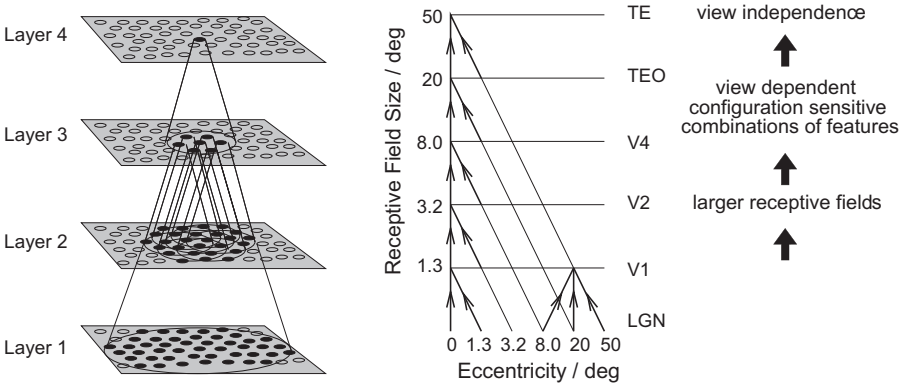

Fig. 3. Left: Stylised image of the four layer network. Convergence through the network is designed to provide fourth layer neurons with information from across the entire input retina. Right: Convergence in the visual system V1: visual cortex area V1; TEO: posterior inferior temporal cortex; TE: inferior temporal cortex (IT).

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

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

object, which is passed through to the first layer of VisNet. The input filters used are computed by weighting the difference of two Gaussians by a third orthogonal Gaussian according to the following:

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

where f is the filter spatial frequency, θ is the filter orientation, and ρ is the sign of the filter, i.e. ±1. Individual filters are tuned to spatial frequency (0.0625–0.5 cycles/pixel); orientation (0–135° in steps of 45°); and sign ($_{-}^{+}1$). The number of layer 1 connections to each spatial frequency filter group is given in Table 2. Past neurophysiological research has shown that models based on difference-of-Gaussians functions are superior to those based on the Gabor function or the second differential of a Gaussian (Hawken & Parker, 1987).

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 δ controls the contrast and σ controls the width, and a and b index the distance away from the centre of the filter

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

The lateral inhibition parameters are given in Table 3.

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

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

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

These are standard VisNet sigmoid parameter values which have been previously optimised to provide reliable and robust performance (Stringer & Rolls, 2008; Stringer et al., 2006, 2007).

Table 3
Lateral inhibition parameters.

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

2.5. Training procedure

There were two different training conditions as follows: (i) the two arms rotating together in lock-step in 40 equal steps, and (ii) the two arms rotating independently in the same fashion. In each simulation one presentation of the full image set containing all configurations of body parts constituted one training epoch.

At each image presentation the activation of individual neurons was calculated, then their firing rates were calculated, and the synaptic weights were updated. In this manner, the network was trained one layer at a time starting with layer 1 and finishing with layer 4. One hundred training epochs were used for each of layers 1–4. The learning rates for layers 1–4 were 0.109, 0.1, 0.1 and 0.1, respectively. The population sparseness of the neuronal firing rates was set to 0.1 for layers 1–4.

3. VisNet simulation results

When the network was tested after being trained on the two arms rotating together in lock step, cells learnt to respond to all of the transforms of both arms. Figs. 4 and 5 show cell response plots for cell (14,17) selected at random in the output layer of VisNet, as the arms were rotating in 40 equal steps. Fig. 4 shows that the cell was firing randomly before training. However after training the cell learnt to fire to all of the transforms of both arms (Fig. 5), thus failing to separate out the two arms from each other. No cells were found that responded to one arm but not the other, suggesting that the network failed to build separate representations of the two arms. These results show that the network failed to form separate representations of the two arms.

When the network was trained on the two arms translating independently from each other, cells learnt to respond invariantly to at least 25% of the views of one arm but not the other. Figs. 6 and 7 show cell response plots for cell (23,10) selected at random in the output layer of VisNet. Fig. 6 shows that the cell was firing randomly before training. However after training the cell learnt to fire to 18 subsequent transforms of the left arm but not to any transforms of the right arm (Fig. 7). A similar cell for the right arm is shown in Fig. 8. Fig. 9 (top) shows a sample of three cells in the output layer which learnt different exclusive partial transform invariant representations of the left arm, whereby these cells did not fire to any transforms of the right arm. Together these cells covered all 40 transforms of the left arm and therefore the firing of any cell in

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.

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

Table 4
Sigmoid parameters.

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

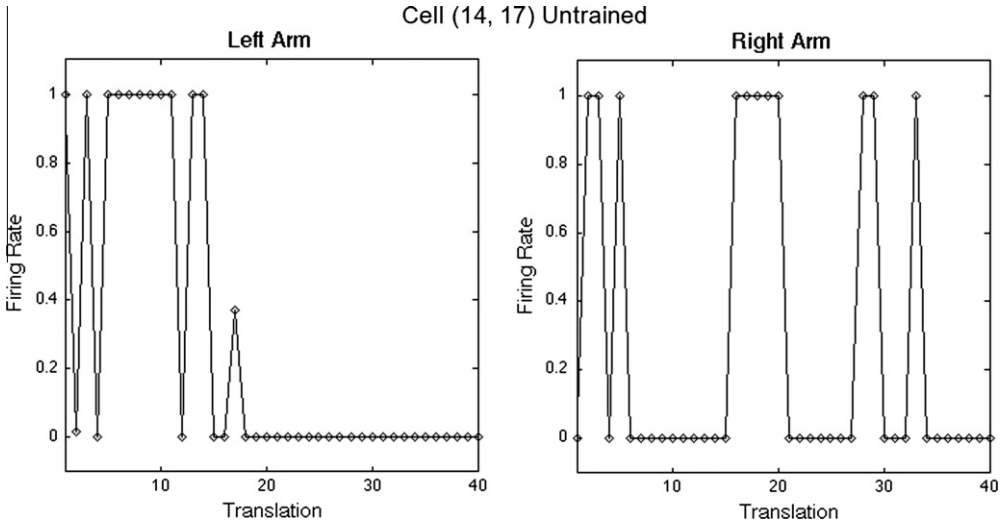

Fig. 4. Simulation results before training. The figure shows the firing rate responses of cell (14,17) in the 4th (output) layer of VisNet to the two arms rotating in 40 equal steps. It can be seen that the cell responds randomly to different transforms of the two arms.

that population could alert the subsequent brain areas that the left arm was present on the retina. Similar cells were found for the right arm (Fig. 9 (bottom)). These results show that the network was able to create separate representations of the two arms with partial (e.g. 25%) transform invariance when independent motion was introduced for the two arms.

Cells were also found that responded to the transforms of both arms, therefore demonstrating that the network was still able to form a representation of the whole object, which in this case consisted of the two arms, as well as its constituent parts. The percentages of the populations of cells that responded to each arm as well as the whole object are presented in Table 5.

4. Discussion

Two simulations have been described in this paper, which supported the hypothesis that independent movement is sufficient in order for the visual system to build separate transform invariant representations of different body parts. It has been shown that when two arms move together, one representation develops, whereby the two arms are seen as one entity. However when independent motion is introduced to the two arms, the visual system is able to build separate representations of them while still being able to form a representation of the whole object. The mechanism employed for invariance learning in this paper was Continuous Transformation (CT) learning. CT learning uses the spatial continuity between the translations of individual objects as they transform in the real world, combined with associative learning of feedforward connection weights. Different parameters have been investigated for the simulations described in this paper, with the number of epochs ranging from 50 to 150, and learning rate ranging from 0.05 to 0.2. The results have been found to be very robust.

Understanding how invariant representations of objects as well as their constituent parts develop in the ventral visual stream without top-down attentional signals is an important question. With inanimate objects, it is plausible to assume that the different constituent parts may be detached from the object and seen in isolation. This might enable the visual system to build representations of these parts. However in living organisms, body parts are not usually seen detached from the rest of the body, and yet separate

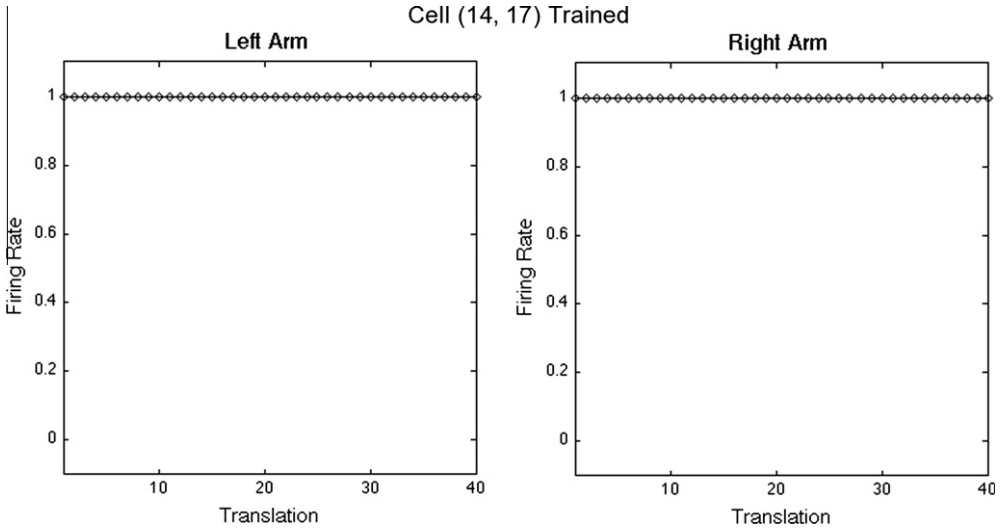

Fig. 5. Simulation results after training with the arms rotating together in lock-step. The figure shows the firing rate responses of cell (14,17) in the 4th (output) layer of VisNet to the two arms rotating in 40 equal steps. It can be seen that the cell responds with transform invariance to both arms.

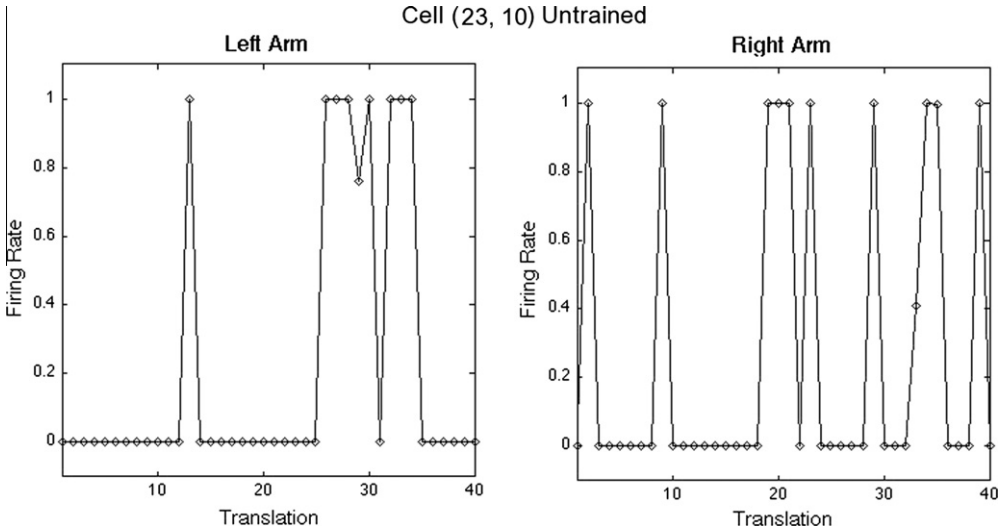

Fig. 6. Simulation results before training. The figure shows the firing rate responses of cell (23, 10) in the 4th (output) layer of VisNet to the two arms rotating in 40 equal steps. It can be seen that the cell responds randomly to different transforms of the two arms.

representations of them exist in the visual system. The current study has shown how these representations might be formed through independent motion of the body parts.

Previous research (Stringer & Rolls, 2008; Stringer et al., 2007) has shown how statistical properties of stimuli can help the ventral visual stream build separate representations of objects when multiple objects are always present during training. The researchers have shown how when features within an object occur together more often than features between objects, separate object representations are formed. Stringer and Rolls (2008) have shown that a stimulus set of at least six objects is necessary in order for this statistical decoupling to happen when the network is trained on all possible pairs of objects. That is, it was necessary to have six objects in the training set in order to have enough different combinations of the objects for the statistical decoupling to happen.

The current study has shown that a similar principle can work when the stimulus set consists of only two objects and they are always shown together during training. Whereas in the study by Stringer and Rolls (2008) the object pairs were rotating together, whereby each rotational view of one object was always presented with the same rotational view of the other object, in the current study independent motion was introduced. During independent motion each view of one object was seen with each view of the other object, thus enabling the statistical decoupling to happen between the transforms of the objects. Features within a view of one object were occurring together more often than features between the view of that object and all the views of the other object. In this way the views of the different objects were separated from each other before being linked together for each object with Continuous Transformation (CT) learning to build transform invariant representations of that object.

The simulations in this study might explain how representations of separate body parts are developed in the brain in the real world. During a lifetime of experiences with living objects whose body parts engage in independent motion, it is expected that all possible combinations of translations and views of the body parts will be seen, therefore providing enough statistical decoupling for representations of the separate body parts to develop in the brain.

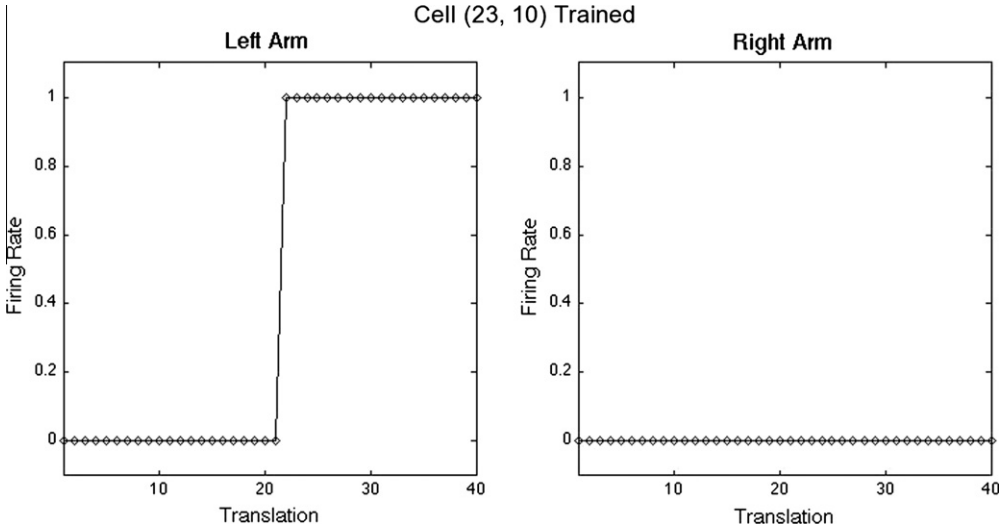

Fig. 7. Simulation results after training with the arms translating independently from each other. The figure shows the firing rate responses of cell (23,10) in the 4th (output) layer of VisNet to the two arms rotating in 40 equal steps. It can be seen that the cell responds with partial transform invariance to the left arm and does not respond to any transform of the right arm.

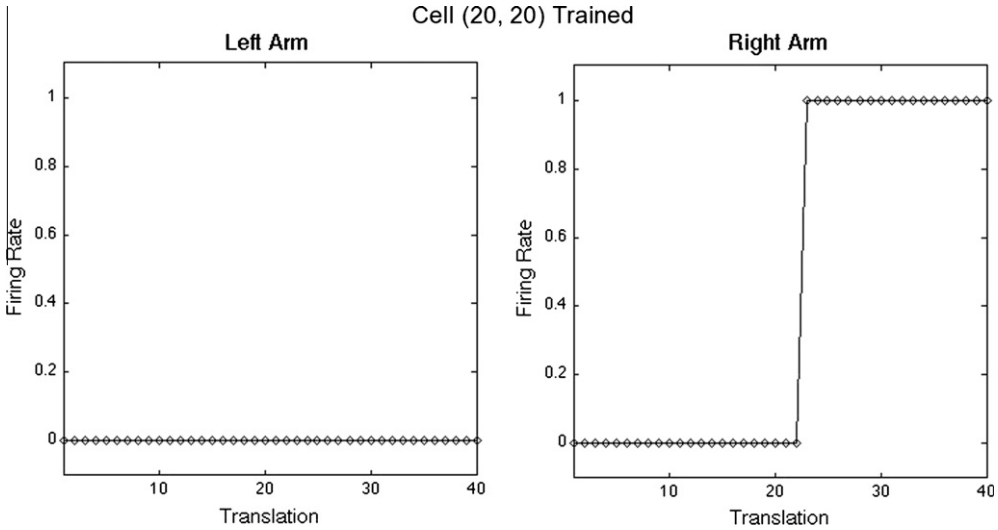

Fig. 8. Simulation results after training with the arms translating independently from each other. The figure shows the firing rate responses of cell (20,20) in the 4th (output) layer of VisNet to the two arms rotating in 40 equal steps. It can be seen that the cell responds with partial transform invariance to the right arm and does not respond to any transform of the left arm.

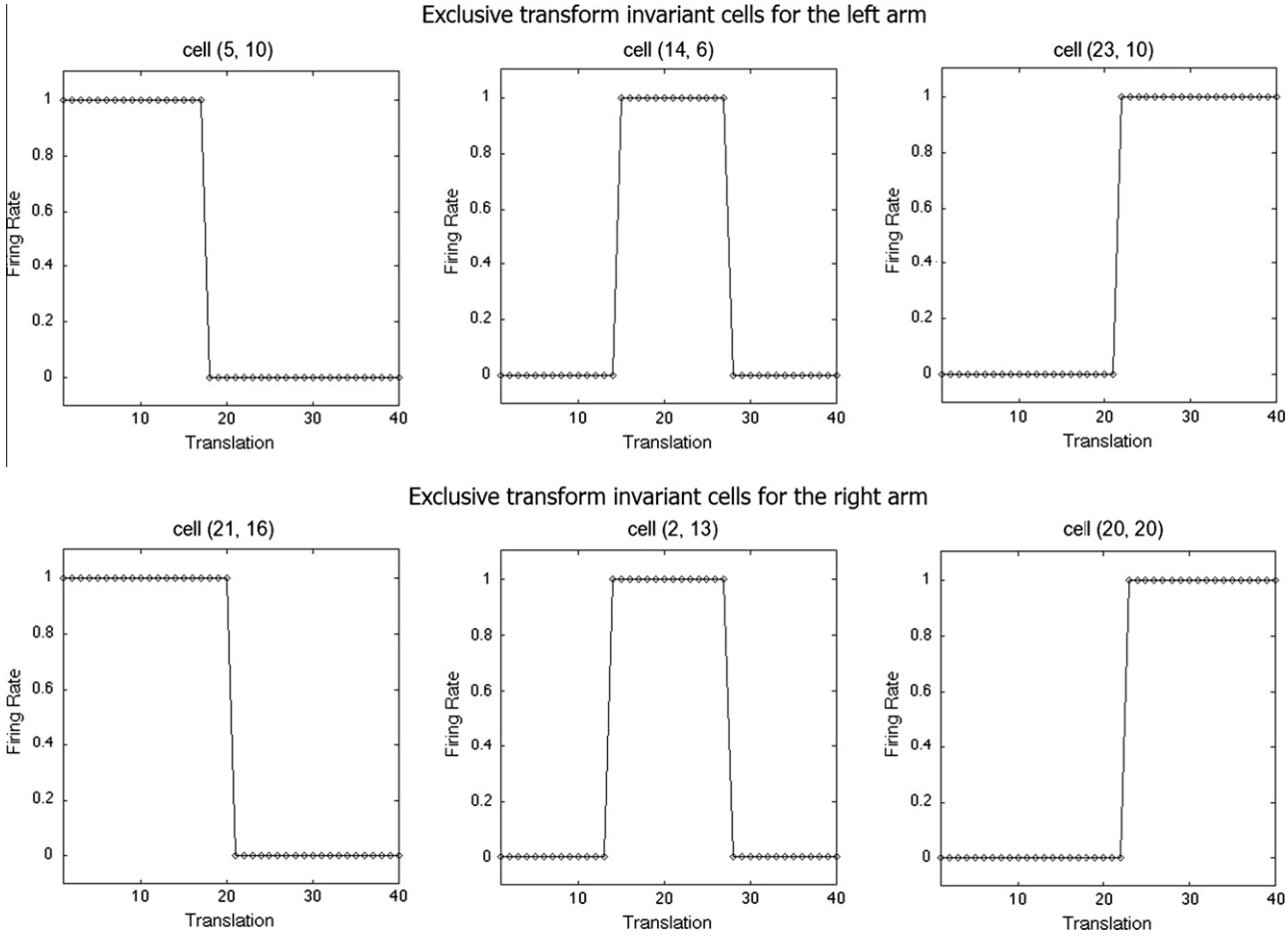

Fig. 9. Simulation results after training with the arms translating independently from each other. Top row: the firing rate responses of three cells in the 4th (output) layer of VisNet to the left arm rotating in 40 equal steps. It can be seen that the cells respond with partial transform invariance to the left arm and between the three cells all 40 transforms of the left arm are represented. These cells are exclusive to the left arm, and they do not respond to any transform of the right arm (not shown). Bottom row: the firing rate responses of three cells in the 4th (output) layer of VisNet to the right arm rotating in 40 equal steps. It can be seen that the cells respond with partial transform invariance to the right arm and between the three cells all 40 transforms of the right arm are represented. These cells are exclusive to the right arm, and they do not respond to any transform of the left arm (not shown).

Table 5
Percentages of populations of cells in the 4th (output) layer of VisNet that responded to each of the two arms as well as the whole object when the arms were rotating in lock step (no independent motion condition) and when independent motion was present.

	Ind. motion (%)	No ind. motion (%)
Right arm	4	0
Left arm	35	0
Whole object	55	100

This is in line with the findings of Chan et al. (2010), who suggested that body representations in the brain depended on life-long experiences and reflected the statistics with which the stimuli occured. The implication of independent motion in visual object segmentation into body parts is also in line with the functional neuroimaging and single-neuron recording studies, which have found that the superior temporal sulcus is implicated in processing biological motion as well as body representations (Grossman et al., 2000; Howard et al., 1996; Jellema et al., 2000; Puce et al., 1998; Wachsmuth et al., 1994).

In the simulations described above, the network was trained on two arms moving independently of each other. We also ran some additional simulations (not shown), in which a static body trunk was added to the rotating arms. We found that the effect of independent motion between the arms can be reduced if a static common body part such as a body trunk is introduced because the body trunk becomes coupled to both of the arms. In this situation some cells responded to both arms, while other cells, which responded selectively to only one arm, had a reduced level of invariance. However when the common body part engages in independent motion, the visual system's performance is restored and it is able to form separate representations of the two arms as well as the body trunk once again.

Even though in the current study a simplified version of the problem has been investigated to test the mechanism of independent motion, whereby schematical representations of the body parts were used and limited 2D rotation was implemented, the same process may work with more complex objects and motions. It is hypothesised that any biological motion is enough to provide sufficient statistical decoupling between the body parts in real life objects. Nevertheless further simulations should be run with more realistic 3D stimuli to investigate whether more life like independent motion can still provide enough statistical decoupling for separate representations of individual body parts to be formed. In order to run these simulations the current model of the ventral visual stream, VisNet, needs to be scaled up to increase the number of neurons in each layer including the retina. This will increase the resolution of the visual input and will ensure that the network is powerful enough to perform well with the more realistic stimuli.

Independent motion is just one principle that might aid the visual system in the process of object segmentation into body parts. It is unable to account for certain types of segmentation, such as the segmentation of facial features for example, since structures such as nose or ears do not engage in independent motion, and yet they are recognised as separate body parts. However even though the independent motion itself cannot account for the segmentation of parts that do not move, the principle of statistical decoupling behind it is hypothesised to still work in these instances. It is hypothesised that in the example of human nose and ears recognition, the exposure to a great number of different faces throughout a lifetime creates the necessary statistical decoupling in order to segment the faces into constituent parts. This is hypothesised to happen because the faces share certain characteristics, in that they all have the same basic structure, but have different individual features. This creates the necessary statistical decoupling between these features across the different faces which means that separate representations of them can be created without the need for independent motion. However further simulations need to be run in order to test this theory.

The results of the simulations described in this paper fit with the theory of Downing et al. (2001), who suggested that bodies were processed by a specialised module in the extrastriate body area in the lateral occipitotemporal cortex, with the help of biological motion, which is processed by the nearby superior temporal sulcus. Although the results of the current study suggest that representations of body parts are built through the same mechanism as representations of other object categories, this study has demonstrated the importance of biological motion in this process.

In this paper, we have not addressed the problem of how particular kinds of biological motion might be recognised by the neural network model because this was not necessary for the network to learn the representations of the separate body parts. The problem of how a network might learn to recognise biological motion has been investigated previously by a number of other authors such as Giese and Poggio (2003) and Casile and Giese (2005). These models are similar to the one used in the current paper in that they are also hierarchical, with increasing feature complexity throughout the hierarchy.

This study has demonstrated how the ventral visual stream can form separate representations of individual body parts when they are always presented together during training. This was shown to be possible through pure bottom-up processes without any top-down attentional influences. It was shown that independent motion is sufficient for this separation to occur. Statistical properties of the stimuli have been shown to be the reason why independent motion works, since it leads to the features within a view of a body part to be seen together more often than the features between that view and the views of other body parts. The current theory is appealing because it shows that living organisms may be segmented into body parts through the interaction between the statistics of the visual input and the architecture of the ventral visual pathway without the need for attentional feedback influences. Therefore the segmentation was an emergent property of the interaction.

References

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*, 115–147.

Casile, A., & Giese, M. A. (2005). Critical features for the recognition of biological motion. *Journal of Vision, 5*, 348–360.

Cave, C. B., & Kosslyn, S. M. (1993). The role of parts and spatial relations in object identification. *Perception, 22*, 229–248.

Chan, A. W. Y., Kravitz, D. J., Truong, S., Arizpe, J., & Baker, C. I. (2010). Cortical representations of bodies and faces are strongest in commonly experienced configurations. *Nature Neuroscience, 13*, 417–418.

De Winter, J., & Wagemans, J. (2006). Segmentation of object outlines into parts: A large-scale integrative study. *Cognition, 99*, 275–325.

Downing, P. E., Chan, A. W., Peelen, M. V., Dodds, C. M., & Kanwisher, N. (2006). Domain specificity in visual cortex. *Cerebral Cortex, 16*, 1453–1461.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science, 293*, 2470–2473.

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

Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience, 4*, 179–192.

Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., et al. (2000). Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience, 12*, 711–720.

Hawken, M. J., & Parker, A. J. (1987). Spatial properties of neurons in the monkey striate cortex. *Proceedings of the Royal Society of London B, 231*, 251–288.

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

Hoffman, D. D., & Richards, W. A. (1984). Parts of recognition. *Cognition, 18*, 65–96.

Hoffman, D. D., & Singh, M. (1997). Salience of visual parts. *Cognition, 63*, 29–78.

Howard, R. J., Brammer, M., Wright, I., Woodruff, P. W., Bullmore, E. T., & Zeki, S. (1996). A direct demonstration of functional specialization within motion-related visual and auditory cortex of the human brain. *Current Biology, 6*, 1015–1019.

Hubel, D. H., & Wiesel, T. N. (1963). Single-cell responses in striate cortex of kittens deprived of vision in one eye. *Journal of Physiology, 26*, 1003–1017.
Jellema, T., Baker, C. I., Wicker, B., & Perrett, D. I. (2000). Neural representation for the perception of the intentionality of actions. *Brain and Cognition, 44*, 280–302.
Kurbat, M. A. (1994). Structural description theories: Is RBC/JIM a general-purpose theory of human entry-level object recognition? *Perception, 23*, 1339–1368.
Lamberts, K., & Freeman, R. P. J. (1999). Building object representations from parts: Tests of a stochastic building object representations from parts: Tests of a stochastic sampling model. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 904–926.
Lamote, C., & Wagemans, J. (1999). Rapid integration of contour fragments: From simple filling-in to parts-based shape description. *Visual Cognition, 6*, 345–361.
Marr, D. (1982). *Vision: A computational investigation into the human presentation and processing of visual information*. San Francisco, CA: W.H. Freeman.
Maurer, D., & Salapatek, P. (1976). Developmental changes in the scanning of faces. *Child Development, 47*, 523–527.
Pasupathy, A., & Connor, C. E. (2001). Shape representation in area V4: Position-specific tuning for boundary conformation. *Journal of Neurophysiology, 86*, 2505–2519.
Pasupathy, A., & Connor, C. E. (2002). Population coding of shape in area V4. *Nature Neuroscience, 5*, 1332–1338.
Perry, G., Rolls, E. T., & Stringer, S. M. (2006). Spatial vs temporal continuity in view invariant visual object recognition learning. *Vision Research, 46*, 3994–4006.
Pinsk, M. A., Arcaro, M., Weiner, K. S., Kalkus, J. F., Inati, S. J., Gross, C. G., et al. (2009). Neural representations of faces and body parts in macaque and human cortex: A comparative FMRI study. *Journal of Neurophysiology, 101*, 2581–2600.
Pinsk, M. A., DeSimone, K., Moore, T., Gross, C. G., & Kastner, S. (2005). Representations of faces and body parts in macaque temporal cortex: A functional MRI study. *Proceedings of the National Academy of Sciences USA, 102*, 6996–7001.
Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *Journal of Neuroscience, 18*, 2188–2199.
Rolls, E. T., & Deco, G. (2002). *Computational neuroscience of vision*. Oxford: Oxford University Press.
Rolls, E. T., & Milward, T. (2000). A model of invariant object recognition in the visual system: Learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Computation, 12*, 2547–2572.
Rolls, E. T., & Stringer, S. M. (2001). Invariant object recognition in the visual system with error correction and temporal difference learning. *Network, 12*, 111–129.
Royer, S., & Parè, D. (2003). Conservation of total synaptic weight through balanced synaptic depression and potentiation. *Nature, 422*, 518–522.
Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition, 80*, 1–46.
Schwarzlose, R. F., Baker, C. I., & Kanwisher, N. (2005). Separate face and body selectivity on the fusiform gyrus. *Journal of Neuroscience, 25*, 11055–11059.
Shelton, J. R., Fouch, E., & Caramazza, A. (1998). The selective sparing of body part knowledge: A case study. *NeuroCase, 4*, 339–351.
Siddiqi, K., Tresness, K. J., & Kimia, B. B. (1996). Parts of visual form: Psychophysical aspects. *Perception, 25*, 399–424.
Singh, M., Seyranian, G. D., & Hoffman, D. D. (1999). Parsing silhouettes: The short-cut rule. *Perception & Psychophysics, 61*, 636–660.
Stringer, S. M., Perry, G., Rolls, E. T., & Proske, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biological Cybernetics, 94*, 128–142.
Stringer, S. M., & Rolls, E. T. (2008). Learning transform invariant object recognition in the visual system with multiple stimuli present during training. *Neural Networks, 21*, 888–903.
Stringer, S. M., Rolls, E. T., & Tromans, J. M. (2007). Invariant object recognition with trace learning and multiple stimuli present during training. *Network, 18*, 161–187.
Van Lier, R., & Wagemans, J. (1998). Effects of physical connectivity on the representational unity of multi-part configurations. *Cognition, 69*, 1–9.
Wachsmuth, E., Oram, M. W., & Perrett, D. I. (1994). Recognition of objects and their component parts: Responses of single units in the temporal cortex of the macaque. *Cerebral Cortex, 4*, 509–522.
Wallis, G., Rolls, E., & Foldiak, P. (1993). Learning invariant responses to the natural transformations of objets. In *Proceedings of the international joint conference on neural networks (IJCNN'93)* (pp. 1087–1090), Nagoya, Japan.
Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology, 51*, 167–194.
Wiskott, L., & Sejnowski, T. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation, 14*, 715–770.