

1 The visually-guided development of facial representations in the primate ventral visual  
2 pathway: a computer modelling study

3 Akihiro Eguchi, Glyn W. Humphreys <sup>†</sup>, and Simon M. Stringer

4 Department of Experimental Psychology, Oxford University, UK

5 Author Note

6 Some of the data and the ideas reported in this article have been presented at  
7 23rd Annual Computational Neuroscience Meeting: CNS2014, Quebec city.

---

<sup>†</sup>Deceased 14 Jan, 2016

## Abstract

Experimental studies have shown that neurons at an intermediate stage of the primate ventral visual pathway, occipital face area, encode individual facial parts such as eyes and nose while neurons in the later stages, middle face patches, are selective to the full face by encoding the spatial relations between facial features. We have performed a computer modeling study to investigate how these cell firing properties may develop through unsupervised visually-guided learning. A hierarchical neural network model of the primate's ventral visual pathway is trained by presenting many randomly generated faces to the network while a local learning rule modifies the strengths of the synaptic connections between neurons in successive layers. After training, the model is found to have developed the experimentally observed cell firing properties. In particular, we have shown how the visual system forms separate representations of facial features such as the eyes, nose and mouth as well as monotonically tuned representations of the spatial relationships between these facial features. We also demonstrated how the primate brain learns to represent facial expression independently of facial identity. Furthermore, based on the simulation results, we propose that neurons encoding different global attributes simply represent different spatial relationships between local features with monotonic tuning curves or particular combinations of these spatial relations.

**Keywords:** primate ventral visual pathway, face processing, facial expression, neural network model, trace learning

## Disclosures

- This research was supported financially by the Oxford Foundation for Theoretical Neuroscience and Artificial Intelligence. The foundation had no other role than providing financial support.
- All authors contributed significantly to the work, and have read and approved the final manuscript.
- The authors declare that they have no conflict of interest.

The visually-guided development of facial representations in the primate ventral visual pathway: a computer modelling study

## Introduction

The ability of the brain to analyse and recognize faces under natural viewing conditions is unmatched by today’s computer vision systems. In order to achieve this singular ability, the primate brain develops and utilizes a rich tapestry of cells that encode different kinds of visual information about faces. For example, some neurons respond to the presence of facial features such as the eyes, nose, or mouth, while other neurons encode the many spatial relationships between these facial features (Freiwald et al., 2009). Some neurons also encode global properties such as facial identity or expression (Morin et al., 2014; Hasselmo et al., 1989). Our ability to process and recognise faces utilises this rich tapestry of different kinds of visual information. Understanding how these diverse visual representations develop through sensory-guided learning may help to inform future research into computer vision for facial analysis and recognition.

## Hierarchical representations of faces along ventral visual pathway

Functional magnetic resonance imaging (fMRI) studies in humans have revealed several cortical regions within the temporal lobe, which are exclusively dedicated to face processing (Perrett et al., 1992; Kanwisher et al., 1997; Pitcher et al., 2011; Zhang et al., 2012). In particular, there is evidence for hierarchical processing. For example, an early stage of processing, the occipital face area (OFA) in the inferior occipital gyrus, has been found to contribute to face perception by responding to individual facial features such as the eyes, nose, and mouth (Pitcher et al., 2011). While a later stage of processing, the fusiform face area (FFA) in the lateral fusiform gyrus, has been found to integrate such information by responding more strongly to intact rather than scrambled faces (Kanwisher et al., 1997; Zhang et al., 2012). Recently, it has also been reported that the face areas may also exhibit “faciotopy” where different cortical patches represent different face features, and the cortical distances between the feature patches

reflect the physical distance between the features in a face (Henriksson et al., 2015).

In macaques, several face sensitive areas have been identified in the temporal lobe, which are known as *face patches* (Gross et al., 1972). It has been argued that the homologue of the FFA in macaques is the *middle face patch* (Tsao et al., 2006). In one single unit recording study by Freiwald et al. (2009), cartoon faces were presented to the monkey. The cartoons were systematically modified by varying the number of facial features present, as well as the spatial relationships between the features such as the distance between the eyes. It was found that the middle face patch also integrated information across facial features. That is, neurons were found to respond to different combinations of facial features (Figure 1) and the spatial relations between them. Furthermore, the tuning profiles of individual neurons that were selective to such spatial relations were typically ramp-shaped between two extremes, which even transgressed the limits of realistic face space (Figure 2). They reported that the responses of neurons encoding spatial relations between facial features were thus amplified for extreme values of these relations compared to intermediate values. Such amplifications may explain the neural mechanisms for the results of experiments showing that faces which are more deviant in appearance are recognized better than those that are more typical (Rhodes, 1997; Benson and Perrett, 1991; Bruce and Young, 2011).

In this paper we investigate through computer modelling how some neurons may learn to respond selectively to individual facial features, or subsets of facial features, even when the model is always trained on whole faces with all of the facial features present. We also investigate how some other neurons learn to encode the spatial relations between facial features, such as distance between the eyes, with monotonic tuning profiles.

## Representations of global attributes of faces

In addition to individual facial features, the primate visual system is able to process various global attributes of faces such as identity, emotional expression, age, race, gender, etc. (Homola et al., 2012; Freeman et al., 2010; Morin et al., 2014;

Hasselmo et al., 1989). Past theoretical work has suggested that the different attributes of a face such as its identity and expression are processed by functionally and anatomically separated pathways. For example, a highly influential psychological model of face processing proposed by Bruce and Young (1986) hypothesized a series of distinct stages involved in face processing. Consistent with this model, experimental studies in macaques have reported that distinct sub-populations of neurons encode different global facial attributes across a number of areas of the primate visual system. For example, it has been shown that the inferior temporal gyrus (TE) contains cells that are primarily selective to facial identity, while the adjacent superior temporal sulcus (STS) contains cells that primarily respond to facial expression (Hasselmo et al., 1989; Perrett et al., 1992; Engell and Haxby, 2007; Wegrzyn et al., 2015). Moreover, a recent study has reported that the number of neurons that encode global attributes such as identity and expression increase along the visual pathway (Morin et al., 2014). This is a quite extraordinary finding since primates are usually exposed to whole faces during visual development. The question is then how might the visual system separate the representations of these global facial attributes, which are always seen together, into different brain areas.

Based on such functional and anatomical specialization, Haxby et al. (2000) hypothesized the existence of a ‘core system’ that is dedicated for visual analysis in the temporal lobe. The core system includes the OFA that detects simple features of faces, the STS that processes changeable attributes of faces such as expressions, and the FFA that processes invariant attributes of faces such as identity.

However, these previous theoretical and experimental studies do not explain the precise learning mechanisms by which these neuronal representations of global attributes, such as identity and expression, may become mapped onto separate processing areas in the later stages of the visual system. Recently, Tromans et al. (2011) developed the first neural network model demonstrating how physically separated representations of facial identity and expression may develop through a biologically plausible process of unsupervised competitive learning. Nonetheless, this modelling

study used highly idealised cartoon faces, in which these two global attributes were artificially encoded by different facial features. In the simulations described below, we investigate these learning mechanisms using much more realistic face stimuli produced using the FaceGen 3D face modelling software package, which generates stimuli based on real faces. This permits a fine-grained study of how facial representations gradually develop through successive stages of processing within the network, until different forms of global attribute eventually appear in distinct regions of the highest layers.

## Computational modelling studies

While many modelling studies have investigated various kinds of processing in the primate visual system (Eliasmith et al., 2012; Serre et al., 2005), most of these investigations have not been concerned with uncovering the synaptic learning mechanisms by which the visual representations develop in the first instance. However, there is a large body of experimental evidence for learning of visual form recognition within the temporal lobes (Wallis, 2013). For example, Baker et al. (2002) showed that exposure to abstract shapes formed by combining multiple parts enhanced both parts-level and holistic shape tuning of neurons in the Inferior temporal cortex (IT). Studies with fMRI have also reported large-scale alteration of the organization and selectivity of temporal lobe cortex in humans after training with visual stimuli (Beeck et al., 2006; Gillebert et al., 2008). Additionally, although the discrimination of non-face objects is known to be more difficult than for faces, training on non-face objects improves the discrimination of these stimuli to nearly that of faces (Yue et al., 2006). Thus, how visual representations develop through a biologically plausible process of visually-guided learning is a key question that needs to be addressed by theoreticians, and is a fundamental aspect of the model simulations presented in this paper.

Lades et al. (1993) presented the first self-organising neural model that developed representations of the spatial relationships between facial features. Their model employed a feature based approach to face recognition via active dynamic linking of features (von der Malsburg, 1981; von der Malsburg and Schneider, 1986). The model

uses an input representation in which each face is convolved with a set of Gabor filters across the visual field. The output layer of the network constructs a graph representation of the face, with each node in the graph representing a particular facial feature, and each link representing the feature relation (Lades et al., 1993). The model was proposed to be biologically realistic because the development of the output face representations is unsupervised. However, it is not clear how a population of neurons in the brain may store facial representations in the form of graphs.

A more biologically plausible approach to modelling how transform (e.g. location or view) invariant representations of faces and non-face objects may develop through unsupervised, associative learning mechanisms in the higher stages of the ventral visual pathway was carried out by Wallis and Rolls (1997). The network architecture is shown in Figure 3. The inputs are represented as columns of V1-like spatial filter activation values similar to the model proposed by Lades et al. (1993). The architecture consists of four layers of competitive neural networks representing successive visual areas V2, V4, TEO (posterior IT) and TE (anterior IT). During training with visual images of faces and other objects, the feedforward synaptic connections between successive layers were modified by a biologically plausible, local, associative learning rule. The study showed that competitive learning allows neurons in the intermediate layers of the model to learn to respond to particular combinations of simple visual features present in faces and non-face objects. By building on these intermediate layer representations, the higher layers were then able to develop transform invariant representations of whole faces. More recently, Wallis (2013) has started exploring various aspects of recognition which are generally regarded as unique to faces such as holistic processing (Tanaka and Farah, 1993), configural processing (Leder and Bruce, 1998), sensitivity to inversion (Yin, 1969; Maurer et al., 2002), the other-race effects (Chance et al., 1982). Besides, the development of face representations within a more biologically accurate spiking neural network model with spike-timing dependent plasticity (STDP) has been presented by Masquelier and Thorpe (2007).

However, these previous studies have not yet fully explained how these



representations correspond to those observed in single unit recording neurophysiology studies develop through successive layers of the model. In the current work, we investigate how successive layers of VisNet develop representations of individual facial features and the spatial relations between these features as reported in the neurophysiological studies described in the introduction. In particular, we show how these cell firing properties may develop naturally through a biologically plausible process of visually-guided learning when the network is trained on realistic face images generated using the FaceGen 3D face modelling software.

## Theory

In this present study, we consider (1) how some neurons along the successive stages of processing learn to represent individual facial features such as the eyes, nose, and mouth given that the visual system is always exposed to whole faces, (2) how some neurons learn to represent particular spatial relationships between facial features, such as the distance between the eyes, with monotonic tuning curves, (3) how some neurons in later stages learn to respond to global attributes such as either a particular identity or expression, and (4) what is the relationship between spatial configurations of facial parts and global representations of face identity and expression.

### **How some neurons learn to represent individual facial features such as the eyes, nose, and mouth**

**The representation of individual local facial features.** Eguchi et al. (2015) considered how neurons in V4 learn to respond selectively to the shape and location of localised boundary contour elements in the frame of reference of the object, and how neurons in areas TEO and TE learn to respond to localised combinations of boundary contour elements. They provided a biologically plausible solution for the development of such cells by showing that the statistical decoupling (Stringer et al., 2007; Stringer and Rolls, 2008) which occurs between different forms of boundary contour element over a large population of different object shapes is a sufficient mechanism for the process. We

hypothesise that a similar learning mechanism may operate to enable the network to learn to represent the individual face parts within a whole face as shown in Figure 4.

Let us assume that each face is comprised of  $n$  different kinds of local facial feature such as the eyes, mouth, and facial outline, and that each such facial feature may occur in  $p$  different possible shapes. In this context, presenting a whole face to VisNet can be seen as presenting an  $n$ -tuple of different facial features simultaneously to VisNet. With  $p$  possible shapes for each facial feature, the number of distinct whole faces that may be constructed is  $p^n$ . This means that if the number of identifiable facial features  $n$  is constant, the number of possible whole face input patterns grows polynomially with  $p$ . This polynomial increase in the representational burden makes it increasingly difficult for the network to develop non-overlapping output representations of all of the faces which are comprised of unique combinations of  $n$  facial features. Therefore, we hypothesise that at a certain point, it becomes less likely that neurons represent all the possible  $p^n$  whole faces, consisting of  $n$ -tuples of features, but rather the neurons may start to represent individual facial features.

For example, consider the case shown in Figure 5. This figure shows a set of faces which are systematically composed of the combination of two possible shapes of the eyes, mouth, and facial outline. Therefore, there are  $n = 3$  facial features and  $p = 2$  shapes of each facial feature, which may be used to construct a total of  $2^3 = 8$  different faces. In the VisNet simulations reported in the first half of the study 1b below, we compared the cell firing properties of the trained network between two conditions: the network trained with only one face ( $n = 3$  and  $p = 1$ ) and the network trained with 8 faces ( $n = 3$  and  $p = 2$ ). In order to minimize the number of cells that happen to exclusively respond to a particular element due to the topologically distributed feed-forward connections of VisNet, we shifted each face across four different retinal locations during training. This would help to confirm the role of statistical decoupling between facial features of different shape, through exposure to many different faces comprised of different combinations of feature shapes, in the development of representations of individual facial features.

In order to test the hypothesis, we recorded the responses of neurons in VisNet to stimuli that contained just one of the facial features used during training. An example set of such test stimuli is shown in Figure 6. These test stimuli allowed us to test whether neurons learned to respond to a specific shape of a particular facial feature as the number of possible shapes  $p$  is varied.

**Shape invariant representations of local facial features.** However, if we continue to increase the number  $p$  of possible shapes of each facial feature, then the range of possible shapes for each feature will begin to form a continuum of gradually changing shapes. Each facial feature will then change its shape in a gradual and continuous manner across different faces. In this situation, the invariance learning mechanism known as continuous transformation (CT) learning (Stringer et al., 2006) may begin to operate. CT learning uses associative learning in competitive networks to build invariant representations by binding together smoothly varying input patterns onto the same output neurons. If an individual facial feature is seen by the network in a large number  $p$  of gradually changing shapes, then the different feature shapes may be bound together onto the same shape invariant neurons in the higher layers of the network by CT learning. This will lead to the development of shape invariant representations of local facial features.

The concept of this learning process is somewhat analogous to that demonstrated in a previous simulation study (Tromans et al., 2012), which investigated the development of transform (view) invariant representations of individual rotating objects when multiple objects were presented rotating together during training. In this simulation study, the objects were presented rotating smoothly across many different views with only small (i.e. 1 degree) changes in orientation between successive transforms. It was found that if two objects were rotated independently of each other, leading to a statistical decoupling between any two particular views of the two objects, then the output neurons in VisNet learned to respond with transform (view) invariance to either one object or the other. The object specificity of the neuronal responses was driven by the statistical decoupling between any two particular views of the two objects

during training, while the view invariance of the responses was driven by CT learning across the gradually changing views of each object. In this way, the network developed separate transform (view) invariant representations for each object.

In our setting, each of the facial features can be regarded as a different object, and the many possible shapes of each facial feature may be regarded as a near continuous space of gradually changing transforms. Because the changing shapes of any two different facial features are independent of each other across many faces, it is expected that the network will first develop neurons that respond to specific local facial features even when the network is always exposed to whole faces during training. However, as we continue to increase the number  $p$  of possible shapes of each facial feature, then we also hypothesize that CT learning will begin to drive the development of shape invariant responses to specific facial features.

In the VisNet simulations reported in the later half of the study 1b below, we only varied the shape of the eyes and tested how shape invariant eye selective cells may develop. In particular, we tested the effects of CT learning on the nature of the facial feature representations that developed in the network during training by varying the number of shapes of the eyes  $p$  (5, 10, or 30 shapes) as shown in Figure 7. We hypothesized that as we increase  $p$ , the facial feature representation turns from shape specific to shape invariant. In order to test the hypothesis, we recorded the responses of neurons in VisNet to stimuli that contained just eyes. An example set of such test stimuli is shown in Figure 8. These test stimuli allowed us to test whether individual neurons learned to respond to just one or a number of different eye shapes as the number of shapes  $p$  was increased.

As a result of the development of representations of individual facial features within the visual hierarchy, we conjectured that neurons in higher layers would start to process these representations and consequently develop various other related response properties. In particular, representations of individual shapes of local facial features could contribute to the development of representations of the spatial relationships between these facial features (Freiwald et al., 2009) as well as the global properties of

faces such as identity and expression (Hasselmo et al., 1989; Morin et al., 2014). At the same time, a collection of shape invariant representations of individual facial features may contribute to the development of global representations of whole faces.

### **How some neurons learn to represent particular spatial relationships between facial features with monotonic tuning curves**

In neurophysiology studies, neurons in the primate middle face patch have been found to encode spatial relationships between facial features, such as the distance between the eyes, with monotonic tuning curves (Freiwald et al., 2009). For example, some neurons that encode the distance between the eyes respond maximally when the eyes are furthest apart, and reduce their responses monotonically as the eyes get closer together. On the other hand, other neurons respond maximally when the eyes are closest together and reduce their responses monotonically as the eyes move further apart. We hypothesise that these monotonic tuning curves develop naturally as a result of competitive learning on the afferent connections into that cortical area when individual neurons receive connections from a physically localised region of the preceding area.

A basic competitive neural network architecture is shown in Figure 9. It consists of a layer of input neurons that send associatively modifiable synaptic connections to a layer of output neurons. The neurons in the output layer compete with each other through inhibitory interneurons to respond to incoming input patterns. Let us identify the neurons in the output layer of this model with a localised sub-population of neurons in the middle face patch, while the input layer contains a localised sub-population of neurons in a preceding cortical area. During learning, the afferent connections to the output neurons are modified by some form of associative learning. The synaptic weight vectors of individual output neurons are also bounded by some form of continual rescaling such as renormalisation, as has been reported in neurophysiology studies (Royer and Pare, 2003). The modification of the afferent connections to the output neurons drives the development of their response properties. These are standard

elements of a competitive neural network architecture (Rolls and Treves, 1998). The question is how neurons in the output layer might learn to represent the spatial relationships between facial features with monotonic tuning curves.

Real neurons in the visual cortex of the brain receive afferent connections from a topologically localised region of the preceding cortical layer. Consequently, local sub-populations of neurons within an area such as the middle face patch may receive afferent connections from input neurons representing only a part of a particular feature space such as the distance between the eyes. In this case, the aforementioned middle face patch neurons may receive a full input representation when the eyes are an intermediate distance apart, but only a partial input representation when the eyes are either far apart or very close together. We hypothesised that this boundary effect at the extrema of the feature space may drive the development of monotonic tuning curves in the middle face patch neurons. Let us illustrate the argument in the context of the competitive network architecture shown in Figure 9 as follows.

Consider the situation where a localised sub-population of output neurons in the middle face patch receives afferent connections from a localised sub-population of input neurons within a preceding cortical area, as shown in Figure 9. Assume that the given input layer neurons represent part of a finite 1-dimensional feature space such as the distance between the eyes. In this idealised example, let the position in the space, i.e. the distance between the eyes, be represented by the position of a localised (e.g. Gaussian) packet of neural activity within the input layer, as shown by the green curve in Figure 9.

The key aspect of this architecture that drives the development of monotonic tuning curves in the output layer is what happens at the two boundaries of the space, when the eyes are either furthest apart or closest together. Specifically, if at each boundary only part (e.g. half) of the input packet is represented, as shown by the red curve in Figure 9, then the output neurons that learn to respond to these particular end locations will end up with relatively large synaptic weights. This is due to these output neurons becoming more tightly tuned to a smaller (end) region of the input layer by

(hebbian) associative learning based on co-activity of the input and output neurons, yet with the magnitudes of the synaptic weight vectors of these output neurons still renormalised over this smaller input region.

If the widths of the input activity packets are relatively broad, then when the input packet is shifted to a more central location within the feature space, the same output neurons, which learned to respond to the end locations, continue to win the competition and respond to the more central locations of the space as well. However, as the input packet shifts away from the ends of the feature space, the responses of these neurons will decline monotonically. Different sub-populations of neurons will learn to respond to the two ends of the feature space, with each sub-population reducing its responses monotonically as the input packet shifts away from its preferred end location.

However, we also hypothesizes that the output neurons only develop monotonic tuning curves if the packet of activity in the input layer is wide enough; otherwise, the end effect breaks down and output neurons develop peaked (e.g. Gaussian) tuning curves. Moreover, if the input space is circular with no end effects, then the output neurons should not develop monotonic tuning curves at all.

The VisNet model architecture shown in Figure 3 is designed to mimic these key aspects of cortical architecture that are needed for the development of monotonic tuning curves. The model is comprised of four competitive layers of neurons. Neurons within each layer receive afferent synaptic connections from a topologically corresponding, localised region of the preceding layer. The synaptic weights may be updated by associative (hebbian) learning rules, with the weight vectors of individual neurons continually renormalised. It was therefore expected that when VisNet is trained on many realistic faces, neurons in the higher layers of model would learn to encode spatial relationships between facial features, such as the distance between the eyes, with monotonic tuning curves.

**How some neurons in later stages of visual processing learn to respond to global attributes of faces such as a particular identity or expression**

The primate visual system can process global attributes of faces such as identity, emotional expression, age, race and gender (Homola et al., 2012; Freeman et al., 2010; Morin et al., 2014; Hasselmo et al., 1989). For example, some neurons in the anterior IT (TE) respond selectively to facial identity, while other neurons in the superior temporal sulcus (STS) respond to facial expression (Hasselmo et al., 1989; Perrett et al., 1992). The important question is how such selective cell response properties could develop given that the visual system is always exposed to both facial identity and expression simultaneously during early visually-guided learning and self-organisation in the visual system.

In earlier work carried out by Tromans et al. (2011), it was shown that when VisNet was trained on a large number of faces, the higher layers of the model developed neurons that either responded to the identity of a face regardless of its emotional expression, or responded to the facial expression irrespective of facial identity. The hypothesised learning mechanism was as follows. If VisNet is exposed to many possible combinations of facial identity and expression during training, then any particular identity is seen only rarely coupled with a particular expression. This creates a statistical decoupling between any particular facial identity and expression. This, in turn, forces individual neurons in the higher layers to learn to respond to either a particular identity regardless of expression, or particular expression regardless of identity. This was demonstrated successfully when VisNet was trained on a matrix of cartoon images of faces with varying identity and expression.

However, this earlier study used a highly idealised set of cartoon faces with two especially unrealistic properties. First, facial identity and expression were represented by different facial features, and thus had non-overlapping representations on the input layer. In particular, facial identity was represented by variation in the shape of the eyes and nose, while facial expression was represented by changes in the shape of the eyebrows and mouth. Thus, the representations of identity and expression were



non-overlapping on the input retina, which is not realistic. With real faces, features such as the eyebrows, eyes, nose, and mouth will all contribute to the representations of both facial identity and expression in a more complex, distributed manner. Accordingly, in this paper, we have investigated whether VisNet will still form neurons that respond to either facial identity or expression even when the network is trained on more realistic faces created using FaceGen, where all of the facial features are involved in representing both facial identity and expression. We hypothesize that this can occur because it is a standard property of competitive networks that under the right conditions, they are able to develop separate (orthogonalised) output representations of distributed (overlapping) input patterns (Rolls and Treves, 1998). These learning mechanisms are demonstrated in simulations presented in the study 3a below.

Secondly, in the study carried out by Tromans et al. (2011), VisNet was trained on every possible combination of facial identity and expression in order to ensure the strongest possible statistical decoupling between these two facial dimensions. This helped to force individual neurons in the higher competitive layers to learn to respond to either a particular identity or expression. However, with real life situations, we do not need to be trained on every possible combination of facial identity and expression in order to learn to recognise these two different facial attributes.

In fact, Tromans (2012) trained VisNet on realistic faces generated using a software FaceGen, the network failed to develop separate representations of facial identity and expression. We hypothesize that this failure was due to the the network being trained on a very dense set of different facial identities and expressions, with the both identity and expression varying almost continuously across their respective dimensions. This rather unnatural set of training faces may have increased the difficulty of neurons in the higher layers developing separate representations of facial identity or expression. In particular, an invariance learning mechanism known as Continuous Transformation (CT) learning (Stringer et al., 2006) may have caused individual neurons in higher layers to learn to respond simply to a large number of gradually changing faces. This is because CT learning is able to bind together smoothly varying

input patterns, such as gradually changing faces, onto the same post-synaptic output neuron. In this way, CT learning may have dramatically reduced the selectivity of neurons for particular facial identities or expressions in the study of Tromans (2012).

We propose that this problem can be remedied by training VisNet on a more realistic, reduced set of face images, with only a limited number of different combinations of facial identity and expression chosen randomly during training. This ensures that the training stimuli do not cover a near continuum of every possible facial identity and expression, which should prevent CT learning from operating. This reduced set of training faces is more realistic than that used by Tromans et al. (2011) since real faces do not actually morph between each other very gradually. In the simulations reported in the study 3a below, training VisNet on the reduced set of realistic faces successfully led to the development of neurons that responded selectively to either facial identity or expression.

The mechanisms underpinning the above hypothesis, that competitive learning can map distributed (overlapping) input patterns to separate (orthogonal) output patterns, may be further elucidated by considering the operation of a simplified competitive network comprised of an input layer that sends associatively modifiable synaptic connections to an output layer. Let us assume that each output neuron receives input from two distinct populations of input neurons, A and B, as shown in Figure 10. Each input population represents a different finite 1-dimensional bounded feature space, such as identity or expression, where the location in the feature space is encoded by the position of a Gaussian packet of activity. During training, Gaussian activity packets are shifted through both of the input populations simultaneously.

According to the basic principle of statistical decoupling, we should see the following effects during learning. If the activity patterns in the two input populations transform in lockstep, so that each location in A is paired with the same location in B, then the output neurons should fail to develop separate representations of the two input spaces. However, if the activity patterns in the two input populations vary independently of each other, so that each location in A is paired with many different

random locations in B etc, then the output neurons should develop separate representations of the two input spaces.

In simulations described in the study 3b below, the effects of statistical decoupling were initially demonstrated for the case, where the representations of the two input spaces are perfectly orthogonal to each other. That is, the two input populations A and B have no cells in common. This situation is analogous to the past study with cartoon faces where identities and expressions were represented orthogonally by different facial features (Tromans et al., 2011). In this simple case, as expected, independent movement of the activity packets in the two input populations leads to the development of separate representations of the two feature spaces in the output layer.

Then we tested whether the same effect is seen when the two input populations are overlapping, i.e. share a number of neurons in common. We first tried 50% overlap between the two input populations, and then tried 100% overlap where each input neuron is a member of both populations A and B. In both cases, individual output neurons learned to respond to either the A population or B population as long as the activity packets in the two spaces moved independently during training. This result shows how competitive learning can form separate representations of two feature spaces, such as facial identity and expression, even when these two dimensions are represented by a common set of input neurons.

We hypothesise that a similar effect will be seen when the VisNet architecture, which consists of a hierarchy of competitive layers, is trained on realistic FaceGen faces, in which both facial identity and expression are represented in a distributed manner by all of the facial features such as the eyebrows, eyes, nose, and mouth.

**What is the connection between neurons representing spatial relationships between facial features and neurons representing global attributes such as facial identity and expression?**

Above, we have discussed how some neurophysiology studies have reported the existence of neurons that represent spatial relationships between facial features such as

the inter-eye distance or height of the eyes (Freiwald et al., 2009), while other studies have reported neurons that encode global attributes of faces such as facial identity or expression (Hasselmo et al., 1989; Perrett et al., 1992; Morin et al., 2014). It is reasonable to expect that the representations of global attributes are dependent upon different spatial configurations of facial parts, with different global attributes such as identity and expression influenced by different spatial configurations of facial parts. We now hypothesise that the cells that encode spatial relationships between facial features in fact largely overlap with the cells representing global facial attributes such as identity and emotion. That is, cells with responses that are correlated with particular global attributes of faces, such as a specific identity or expression, may actually be tuned to a particular spatial relationship between certain facial features that are indicative of that global attribute. For example, the responses of cells that are tuned to a specific shape of the mouth might be also correlated with a particular facial expression such as happy. In this situation, the cell might be regarded as contributing to representing both the shape of the mouth and facial expression.

It has long been known that the neural representation of faces in the primate visual system is distributed, with individual faces represented by many neurons and individual neurons participating in the representation of many faces (Rolls and Treves, 1998). The question is whether a global attribute, such as a particular identity or expression, is represented by a *random* subset of neurons, or whether individual neurons actually represent specific constituent features of the global attribute. In the latter case, a neuron that represents a particular curvature of the mouth might participate in the representation of a happy expression across a number of different facial identities. This is what we are proposing. In this case, the activity of neurons encoding a particular spatial relationship will also be correlated with a particular corresponding global attribute, and may be thought of as participating in the representation of that global attribute.

It is then possible that individual neurons in even higher layers learn to respond to specific combinations of neurons representing the spatial relationships between facial

features that are correlated with a particular global attribute. These higher layer neurons would be tuned to a combination of all the spatial relationships comprising a particular global attribute, and so might be regarded as providing the most abstracted representation of the global attribute, that is, in a way that does not depend on the presence of any one particular spatial relationship between facial features.

### Simulation Studies

In this paper, we carried out three simulation studies using an established hierarchical neural network model of the primate ventral visual pathway, VisNet, which was originally developed by Wallis and Rolls (1997). The standard network architecture is shown in Figure 3. It is based on the following: (i) A series of hierarchical competitive layers with local graded short-range lateral excitation and long-range lateral 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 (Freeman and Simoncelli, 2011; Pasupathy, 2006; Pettet and Gilbert, 1992). (iii) Synaptic plasticity based on a biologically-plausible local learning rule such as the Hebb rule.

In the present simulations, all the visual inputs were pre-processed by a set of Gabor filters that accord with the general tuning profiles of simple cells in V1 (Jones and Palmer, 1987; Cumming and Parker, 1999; Lades et al., 1993). The VisNet architecture used in these simulations consisted of four Self-Organising Maps (SOM) (Kohonen, 1982). The parameters used in the VisNet simulations in this paper are shown in Table 1(a). The gradual increase in the receptive field of cells in successive layers reflects the known physiology of the primate ventral visual pathway (Freeman and Simoncelli, 2011; Pasupathy, 2006; Pettet and Gilbert, 1992; Wallis and Rolls, 1997), and the other parameters were selected based on those that previously optimised performance (Tromans et al., 2011; Rolls and Milward, 2000). Full details of the VisNet architecture are provided in Appendix A.

The VisNet model was trained on realistic images of faces with different identities

and expressions generated using the FaceGen face modelling software. FaceGen builds artificial 3D face images from templates taken from 273 high resolution 3D face scans. The images are averaged, and PCA is used to extract a set of variances from the mean representing facial features such as shape, colour and gender. This in turn gives a normal distribution from which a random coefficient can be chosen, creating a random, realistic face based on a range of alterable features. After training, we investigated whether the neurons in the higher layers of the network had developed response characteristics similar to those reported in neurophysiology studies.

In the series of studies conducted in this paper, we tested whether VisNet developed neurons with response characteristics similar to what has been found in neurophysiology experiments. In the first study, we explored how neurons learn to respond to individual local facial features such as the facial outline, eyes, nose, and mouth, as well as specific global combinations of these features. The second study investigated how some neurons learn to represent the spatial relationships between particular facial features, such as the distance between the eyes, with monotonic tuning curves. Finally, in the third study, we explored how some neurons learn to represent the global attributes of either facial identity or facial expression.

However, unless otherwise stated, these VisNet studies were carried out by testing the same trained network. That is, the VisNet model was trained only once at the beginning, and then the same trained network was tested across all three studies for the various cell response properties. There is one exception to this in the first study, where the network is retrained to investigate how increasing the variation in facial features, such as the eyes, nose, and mouth, drives the development of neurons that respond to the individual features.

During the initial training of VisNet, the network was presented with 450 realistic human faces as shown in Figure 11 and 150 non-face objects as shown in Figure 12. The faces were randomly generated with different identities using the commercial software FaceGen, and the expressions of individual faces were also randomly set along a continuous dimension between happy and sad. Non-face objects were retrieved from

Google 3D warehouse. All stimuli were grayscaled and projected onto an input retina that was  $256 \times 256$  pixels in size.

In the second and third studies, we also carried out some complementary simulations with the simplified network model with only one layer of fully connected, associatively modifiable synapses as shown in Figure 9. The network was trained and tested on 1-dimensional Gaussian input patterns, which provided an idealised representation of a 1-dimensional facial feature space such as the distance between the eyes. This abstracted neural network model allowed a more controlled investigation of the hypothesised mechanisms underpinning the development of the cell response characteristics of interest. The parameters used in the simulations in this paper are shown in Table 1(b). Full details of the network architecture are provided in Appendix A. The purpose of these additional simulations was to investigate deeper into the underlying learning mechanisms using a more simplified and controlled setup.

### **Study 1: The neural representation of local facial features and combinations of features**

**Simulation results of VisNet.** Freiwald et al. (2009) showed that cells in the middle face patch of the primate visual system responded selectively to individual facial features or particular combinations of features. Therefore, in this first study, we investigated the neural representation of individual local facial features, such as the facial outline, eyes, nose, and mouth, as well as global combinations of these features, throughout the hierarchical architecture of VisNet. Specifically, it was explored whether such neuronal responses had developed in VisNet during the initial training on 450 realistic human faces as shown in Figure 11 and 150 non-face objects as shown in Figure 12.

The neurophysiology study carried out by Freiwald et al. (2009) showed that cells in the middle face patch were tuned to different combinations of facial features. For example, some neurons were tuned to the presence of only one particular facial feature, while other cells were tuned to a particular combination of either 2, 3, or 4 facial

features. Therefore, to investigate whether similar cells had developed in VisNet, we tested the network on face stimuli that were comprised of all possible combinations of the four facial features: facial outline, eyes, nose, and mouth. For each possible combination of facial features, we created 15 different facial identities in order to test for generalisation across different facial identities. A subset of the face stimuli used for testing the network is shown in Figure 13.

Similar to the results reported by Freiwald et al. (2009), we found that neurons learned to respond to different combinations of the facial features. Figure 14 shows the firing rate responses of five different 4th layer neurons to face stimuli constructed from different combinations of facial features for 15 distinct facial identities. For example, the first cell (113,1) shown in the figure is more likely to be activated when the facial outline is present. The second cell (82,67) responds strongly whenever the mouth is present. The third cell (80,61) responds when the facial outline and eyes are presented together. The fourth cell (102,62) responds most strongly when the facial outline is present but the mouth is absent. The fifth cell (99,38) responds most when the facial outline and mouth are absent. Similar cell selectivities were also found by Freiwald et al. (2009).

Figure 15 shows the number of 4th layer neurons that responded significantly more strongly ( $P < 0.005$ ) to the presence or absence of a particular number (1, 2, 3, or 4) of facial features before and after training based on paired t-test over identities. The results confirm that, after training, different neurons responded maximally to different numbers of the facial features. Some cells were tuned to only a single facial feature, while other cells responded most to either 2, 3, or 4 facial features.

In order to further quantify the selectivity of neurons to individual face parts in the successive layers, single cell information analysis was conducted as described in Appendix B. Figure 18 shows the single cell information plots for each layer of VisNet in which the testing stimuli are four different face parts (mouth, nose, eyes, and outline) for 15 distinct facial identities shown in Figure 13. The number of cells that reached maximum single cell information is small in the first layer, is the largest in the second and the third layers, but then declines slightly in the fourth layer. These simulation



results reflect the hierarchical representation of faces in the primate visual system discussed in the introduction. Specifically, the simulation results mirror how the occipital face area (OFA) in an early stage of processing learns to respond to individual facial features, while the fusiform face area (FFA) in a later stage of processing subsequently integrates this information.

In addition, we have mapped the 4th layer cells that carry highest single cell information for each facial feature to explore whether “facitopy” has been developed in our simulation as reported in the fMRI study of Henriksson et al. (2015). In their study, the cortical representations of facial features such as the eyes, nose and mouth were found to be arranged in a map that corresponded to their relative positions within the face. The contour plots shown in Figure 17 indicate each sub-region comprised of the top 500 cells that carry the highest single cell information for one of the four facial features: mouth (red), nose (green), eyes (blue), and outline (black). Consistent with the facitopy hypothesis, the distribution of the sub-regions are found to be roughly corresponding to the physical configurations of facial features within the face.

The above results show that along the hierarchy, the network develops separate representations of the local facial features such as the facial outline, eyes, nose, and mouth. However, the question is how the network learns to represent the individual facial features when the network is always exposed to complete faces comprised of all the facial features presented together during training. Since we have identified that the number of face feature selective cells is greater in the intermediate layers (i.e. 2 and 3) than in the output (4th) layer, in the following subsection we focus our analysis of the learning mechanisms underpinning the development of feature selective neurons on the third layer of the network.

**How the network learns to represent individual facial features through competitive learning driven by statistical decoupling between the features.**

*Shape selective facial feature representations.* In the theory section above, we hypothesised that some neurons would become tuned to particular facial features due to the statistical decoupling between any two of these features as the

number of shape variations increases. Specifically, eyes of a particular shape and a particular shaped mouth would be seen together only rarely. This creates a statistical decoupling between these two particular features, which in turn makes it difficult for neurons to learn to respond to this particular combination. In order to carry out a controlled test of this hypothesis, we ran two simulations in which VisNet was trained on faces with  $n = 3$  variable facial features: eyes, mouth, and facial outline. In the simulations, the number of shape variations of each of these facial features  $p$  was set to either 1 or 2. Accordingly, the number of distinct shapes of facial features to be learned is 3 and 6 ( $n \times p$ ), and the number of whole faces presented during training is 1 and 8 ( $p^{(n)}$ ), respectively. Additionally, in order to eliminate the cells that happen to exhibit facial feature selectivity due to the topologically distributed feedforward synaptic connections in the model, each face was presented in a  $2 \times 2$  grid of 4 different retinal locations, which were separated by horizontal and vertical shifts of 10 pixels. For each simulation, after training using the temporal trace learning rule as described in equations (7) and (8) in Appendix A, the network was tested with the set of face stimuli constructed by extracting just one of the three facial features as shown in Figure 6 for  $p = 2$ .

In order to quantify the performance, single cell information analysis was conducted as described in Appendix B. Figure 18 shows normalized single cell information plots for two simulations in which the training stimuli were constructed with  $p$  set to either 1 or 2 shapes for each of  $n = 3$  facial features, eyes, mouth, and outline. Each plot shows the information carried by all of the 3rd layer neurons about a specific shape of one of the three facial features, where the neurons are plotted in rank order along the abscissa. The maximum amount of information possible for the simulations is  $\log_2(n \times p)$ , that is 1.6 or 2.6 bits for  $p = 1$  or 2 respectively. The result shows that the number of cells that learned to carry maximum single cell information increased as  $p$  was increased from 1 to 2. Thus, for the higher value of  $p = 2$ , neurons learned to be more selectively tuned to a specific facial feature shape, which was due to statistical decoupling between different shaped features across multiple faces.

*Shape invariant facial feature representations.* At the same time, we hypothesized that as  $p$  increases, then CT learning will begin to bind together the different shapes of a particular facial feature leading to different subset of neurons that respond to all possible shapes of that feature. In particular, if the network is exposed to many different shapes of eyes covering a near continuum of gradually changing eyes, then the continuous transformation (CT) learning mechanism may bind together the different shapes of eyes onto the same subset of output cells, which would then respond to all eyes. Something similar would occur for the other facial features such as the facial outline and mouth. The end result of these learning mechanisms should be that as the number of shape variations  $p$  for each facial feature increases, more cells should learn to respond selectively to all possible shape variations of just one particular facial feature.

In order to confirm our hypothesis that CT learning was beginning to bind together the shape variations of each particular facial feature as  $p$  increased, we conducted another series of simulations. For each of three further simulations, the network was exposed to larger numbers of faces where the shape of eyes was varied over 5, 10, or 30 shapes during training as shown in Figure 7. Again, in order to eliminate the cells that happen to be exclusively responding to a particular facial feature due to the topologically distributed feed-forward synaptic connectivities, the faces were shifted across four different retinal locations during learning. After training using the temporal trace learning rule (equation 7 and 8 in Appendix A), the network was tested with 50 eyes which were extracted from randomly generated faces as shown in Figure 8. Figure 19 shows the distribution of the number of cells that respond to different numbers of the shapes of eyes. The result when the network was trained with 5 faces is plotted with a dotted line, the results with 10 faces is plotted with a dashed line, and the results with 30 faces is plotted with a solid line. It can be seen that over the three simulations, for larger values of  $p$ , the number of cells that have learned to respond to most of the shape variations of eyes increases. This confirms that as the number of shape variations of a particular facial feature increases, CT learning binds together these shape variations to produce neurons that respond selectively to one particular

facial feature over all possible shapes.

In conclusion, the above simulations show how the network is able to develop neurons that respond to just one particular shape of a facial feature, or particular combinations of facial features, through the statistical decoupling that occurs between facial features when the network is presented with many different faces during training. Moreover, we have shown how some neurons may learn to respond invariantly to all the different shape variations of a particular facial feature through continuous transformation (CT) learning when the number of feature shape variations across different faces is large.

Given these representations of individual facial features within the visual hierarchy, we conjectured that neurons in higher layers would start to process these representations and consequently develop various other related response properties. In particular, representations of individual shapes of local facial features could contribute to the development of representations of the spatial relationships between these facial features (Freiwald et al., 2009) as well as the global properties of faces such as identity and expression (Hasselmo et al., 1989; Morin et al., 2014). At the same time, a collection of shape invariant representations of individual facial features may contribute to the development of global representations of whole faces.

## **Study 2: The representation of spatial relationships between facial features with monotonic tuning curves**

**Simulation results of VisNet.** Freiwald et al. (2009) showed that some neurons in the middle face patch of the primate visual system encoded the spatial relationships between facial features, such as the distance between the eyes, with monotonic tuning profiles. We, therefore, tested whether such neurons had developed in VisNet during the initial training on 450 realistic human faces as shown in Figure 11 and 150 non-face objects as shown in Figure 12. For this purpose, we constructed a set of test face stimuli, in which the geometrical parameters of the facial features were systematically varied to be comparable with the physiological study conducted by

Freiwald et al. (2009). In particular, we varied the dimensions of inter-eye distance, eye-brow angle, eye-height, and mouth shape as shown in Figure 20. For each such dimension of spatial variation, we used faces with five different identities. And for each facial identity, we constructed ten face images by sampling ten different, evenly-spaced feature values of the relevant dimension. The ten selected feature values spanned the entire range of realistic values for that dimension. These face stimuli were presented to VisNet during testing, and the firing rate of each neuron in the network was recorded.

Figure 21 shows eight example neurons (a-h) found in the 4th layer of VisNet which represent different spatial relationships between facial features with monotonic tuning profiles. Neurons a and b encode inter-eye distance, neurons c and d encode eyebrow angle, neurons e and f encode eye height, and neurons g and h encode mouth shape. Visual inspection of the firing rate responses across the 4th layer confirmed that many neurons had developed monotonic tuning responses to variation in these four spatial relationships between facial features.

**Simulation results of the simplified network model with one layer of synapses.** In order to carry out a deeper investigation into the learning mechanisms by which neurons could develop monotonic tuning responses encoding the spatial relationships between facial features, we carried out further simulations in a simplified neural network architecture with one layer of synapses as described in Appendix A and shown in Figure 9. The network was trained and tested on 1-dimensional Gaussian input patterns, which provided an idealised representation of a 1-dimensional facial feature space such as the distance between the eyes. During training, a Gaussian packet of activity is imposed at a series of randomly selected locations on the input layer. At each location of the Gaussian input packet, activity is propagated to the output neurons, and then the synaptic weights are modified using a local associative (hebbian) learning rule with synaptic weight vector normalisation as described in Appendix A. The sigma value that controls the width of the Gaussian input packet,  $\sigma$ , was set to 10 unless otherwise stated. During the testing, the location of the Gaussian packet was moved from neurons 1 to 100 across the input layer, and the firing responses of the

output neurons were recorded for each location. This abstracted neural network model allowed a more controlled investigation of the learning mechanisms responsible for the development of monotonically tuned responses among the output cells.

Figure 22 shows the development of monotonic tuning responses in the simplified network model with one layer of synapses. The figure shows results for three different simulations: (i) network trained with circularly arranged input neurons with wrap-around (top row), (ii) network trained with linearly arranged input neurons with no wrap-around (divisive inhibition) (middle row), and (iii) network trained with linearly arranged input neurons with no wrap-around (combined lateral inhibition and excitation) (bottom row). The columns show the following: (a) the width,  $\sigma$ , of the Gaussian activity packet imposed on the input layer during training and testing, (b) matrix of synaptic weights from input neurons to output neurons, (c) matrix showing activations of output neurons as a Gaussian activity packet is shifted through successive locations on the input layer, and (d) matrix showing firing rates of output neurons as a Gaussian activity packet is shifted through the input layer. The plots show that, regardless of the type of competition implemented, the trained networks with linearly arranged input neurons with no wrap-around (middle and bottom rows) have developed output neurons with monotonic tuning responses to the location of the Gaussian activity packet in the input layer. On the other hand, output neurons did not show monotonic responses in the network trained with circularly arranged input neurons with wrap-around. In such a circular network there are no such end effects on learning, which are needed to drive the development of output neurons with monotonically tuned responses.

Figure 23 shows further results for the three simulations shown in Figure 22. For each of these simulations, Figure 23 shows the behaviour of twelve typical output neurons in separate subplots. In particular, those cells are the cells indexed with 1, 10, 19, 28, 37, 46, 55, 64, 73, 82, 91, and 100 in the output layer. Each subplot shows how the activation and firing rate of the neuron vary as a Gaussian activity packet is shifted through the input layer. Figure 23 confirms that regardless of the type of competition,

the trained network with linearly arranged input neurons with no wrap-around displays output neurons with responses that are monotonically tuned to the location of the Gaussian activity packet in the input layer. Moreover, some output neurons respond maximally when the Gaussian activity packet is presented at the left end of the input feature space, and their responses decline monotonically as the packet is shifted to the right. While other output neurons respond maximally when the Gaussian activity packet is presented at the right end of the input feature space, and their responses decline monotonically as the packet is shifted to the left (Figure 23(b,c)). This demonstrates that the network develops either monotonically increasing or decreasing responses along the feature space as was reported by Freiwald et al. (2009). However, output neurons failed to show monotonic responses in the network trained with circularly arranged input neurons with wrap-around. Instead, the output neurons in the circular network developed peaked responses (Figure 23(a)).

These results are consistent with our original hypothesis described in the theory section above. The key aspect of this network architecture that drives the development of monotonic tuning curves in the output layer is what happens at the two ends of the input space during learning. In the network trained with linearly arranged input neurons with no wrap-around, only part (e.g. half) of the input packet is represented at each end. In this case, the output neurons that learn to respond to the end locations develop relatively large synaptic weights. This is due to these output neurons becoming more tightly tuned to a smaller (end) region of the input layer by associative learning, but with the magnitudes of their synaptic weight vectors still renormalised over this smaller input region. If the widths of the input activity packets are relatively broad, then the same neurons continue to win the competition and respond when the input packet is shifted to a more central location within the input layer. However, as the input packet shifts away from the ends of the input layer, the responses of these neurons will decline monotonically. Different sub-populations of neurons will learn to respond to the two ends of the input layer, with each sub-population reducing its responses monotonically as the input packet shifts away from its preferred end location.

We also explored how varying the standard deviation  $\sigma$  that determine the width of the Gaussian activity packet imposed on the input layer affected the development of monotonic neuronal responses in the output layer. Figures 24 and 25 show the results of four simulations with divisive inhibition, where each simulation used a different value of  $\sigma$  set to 2, 5, 10, and 20, respectively. It can be seen that for a relatively small value of  $\sigma$  equal to 2, the output neurons do not develop monotonic tuning responses (Figure 24 (top row) and 25(a)). However, when  $\sigma$  is increased to 20, then the output neurons do develop monotonically tuned profiles after training (Figure 24 (bottom row) and 25(d)). Simulation results for  $\sigma$  equal to 5 or 10 show intermediate output behaviours. These results show that the output neurons gradually transition to developing monotonic responses as the standard deviation  $\sigma$  that determine the width of the Gaussian input packet increases. Thus, the width of the input packet needs to be reasonably large with respect to the size of the input space in order to drive the development of monotonically tuned output neurons.

The above simulations showed how output neurons may develop monotonic tuning profiles when the network with divisive inhibition is trained with the input activity packet presented across all locations in the input feature space. However, Freiwald et al. (2009) showed that neurons in the middle face patch of the primate visual system maintained their monotonic tuning curves even when the monkey was presented with cartoon faces with unrealistically extreme spatial variations between the facial features, such as unrealistically large inter-eye distances, that could not have been encountered during prior visual experience. Accordingly, our next question was whether our model still develop output neurons that are monotonically tuned over the entire input feature space, including the extremal locations, if the model was trained with the input activity packet presented within only a limited central sub-region of the input feature space. Figures 26 and 27 show the results of three simulations in which the Gaussian activity packet was shifted over different sized central intervals, 25%, 50% and 75%, of the input layer during training. It can be seen that monotonic response curves still develop in the output layer when the input activity packet is presented within only 75 % or 50 % of



the input feature space during training. Both of these two simulations still allow for some degree of truncation of the Gaussian activity packet at the two ends of the input layer, which is required for the development of monotonic tuning profiles in the output layer according to the hypothesis described in the theory section above. These results thus confirm that the training set does not have to cover entire input space in order for the output neurons to develop monotonic tuning curves. This, in turn, offers an explanation for the experimental findings of Freiwald et al. (2009) that neurons in the monkey brain maintain their monotonic tuning curves even when presented with unrealistically extreme spatial variations between the facial features.

Lastly, we ran simulations to test whether the truncation of the Gaussian activity packet at the ends of the input layer during training played a key role in driving the development of monotonically tuned output neurons, as hypothesised in the theory section above. In these simulations of the simplified network, we extended the input layer to include 100 extra neurons on either side of the 100 original input neurons, which gave a total of 300 input neurons. However, during training, the Gaussian activity packet was still presented only within the interval covering the original 100 input neurons. The inclusion of the extra 100 input neurons on either side of the original central region ensured that the Gaussian activity packet was not truncated at the original end locations. This should, according to our hypothesis described in the theory section above, reduce the development of monotonic tuning profiles in the output layer. The result shown in Figure 28 support the hypothesis. In particular, some of the output neurons began to develop non-monotonic peaked (Gaussian) tuning responses as the end effects due to truncated Gaussian input packets broke down as the input layer was extended.

In conclusion, the above simulations show how the network is able to develop neurons that encode spatial relationships between facial features, such as the distance between the eyes, with monotonic tuning curves as reported in physiology (Freiwald et al., 2009). We proposed and provided evidence of the possible developmental mechanism of such cells, which is a result of competitive learning on the afferent

connections into that cortical area when individual neurons receive connections from a physically localised region of the preceding area leading to end effects.

### **Study 3: The representation of global facial attributes such as facial identity and expression**

**Simulation Results of VisNet.** Neurophysiology studies have demonstrated the existence of separate clusters of neurons in the primate visual system that encode either facial identity or facial expression (Hasselmo et al., 1989; Perrett et al., 1992; Morin et al., 2014). The question is how such cell response properties could develop. When Tromans et al. (2011) trained VisNet on cartoon faces of varying identity and expression, the network successfully developed separate clusters of neurons that encoded either facial identity or expression. However, when Tromans (2012) trained VisNet on a continuum of realistic faces generated using FaceGen, the network failed to develop separate representations of facial identity and expression. We hypothesised in the theory section above that this problem can be remedied by training VisNet on a more realistic, reduced set of face images, with only a limited number of different combinations of facial identity and expression. Another limitation of the study of Tromans et al. (2011) was that VisNet was trained on cartoon images of faces in which facial identity and expression were artificially represented by different facial features. We hypothesised in the theory section that VisNet should still form neurons that respond to either facial identity or expression when the network is trained on more realistic faces where facial identity and expression may be represented by common facial features.

We tested whether neurons that responded selectively to either facial identity or expression had developed in VisNet during the initial training on 450 realistic human faces as shown in Figure 11 and 150 non-face objects as shown in Figure 12. For this purpose, we constructed a new test set of realistic face stimuli using FaceGen as follows. We first created a 1-dimensional space of 20 different facial identities, which varied gradually from an extreme Identity A to another extreme Identity B. Then each of these identities was varied over a 1-dimensional space of 20 different expressions from

Sad to Happy. This resulted in a set of 400 face stimuli constructed from 20 identities  $\times$  20 expressions as shown in Figure 29. The trained network was tested on each face stimulus in the set, and the firing-rates of all neurons in the model were recorded.

Figure 30 shows the firing rate responses of typical neurons in the 4th layer of VisNet when tested on the facial stimuli representing combinations of identity and expression shown in Figure 29. Results are shown before and after training on the 450 realistic human faces shown in Figure 11 and 150 non-face objects shown in Figure 12. The individual plots in Figure 30 show how the firing rate of each neuron varies with facial identity and expression. Before training, the neuronal responses do not depend in a structured way on facial identity and expression (Figure 30(a)). However, after training, individual neurons have learned to respond selectively to localised regions of either the space of identities or space of expressions (Figure 30(b)). The first (top) row in Figure 30(b) shows neurons that have learned to respond to expressions near the right of the expression space bounded by Sad, while other neurons in the second row have learned to respond to expressions near the left of the expression space bounded by Happy. In contrast, the third row shows neurons that have learned to respond to identities on the top of the identity space bounded by Identity A, while other neurons in the fourth (bottom) row have learned to respond to identities on the bottom of the identity space bounded by Identity B.

It can be seen that training VisNet has produced neurons that respond selectively to either particular identities or expressions. Furthermore, very interestingly, it can be seen that individual neurons have monotonic responses to the particular global feature dimension, i.e. identity or expression, which the neuron is tuned to. This is reminiscent of the neurons shown above in the study 2, which represent the spatial relationships between facial features with monotonic tuning curves. The question is could there be an underlying connection between these two kinds of neuron. We explore this idea further below. Figure 30 also shows that neurons that represent specific identities tend to be clustered close together, and the same is true for neurons that encode particular expressions. This is due to a combination of short range excitation and long range

inhibition, effecting a self-organising map (SOM), within each layer of VisNet.

Figure 31 shows the results of analysing the amount of single and multiple cell information carried by 4th layer neurons in VisNet about facial identity and expression before and after training (Appendix B). The left column of Figure 31 shows the amount of information about identity conveyed by fourth layer cells. This analysis involved quantising the identity space into five separate contiguous blocks. The maximal amount of information possible in this case is  $\log_2(5) = 2.32$  bits. The right column of Figure 31 shows equivalent results for the amount of information conveyed by fourth layer cells about expression. It can be seen that training has led to a substantial increase in the amount of single and multiple cell information about both facial identity and expression. Thus, information analysis confirms the enhanced selectivity of neurons for either identity or expression after the training. More than 100 neurons carry 1.5 bits or above of single cell information for facial identity, and around 100 cells carry 1 bit or above of single cell information for expression. This is consistent with the monotonic tuning curves shown in Figure 30. Such neurons respond to a localised region at one end of their preferred feature space, e.g. responding to Happy faces but not Sad faces. Such neuronal responses will carry at least 1 bit or more of information about the neuron’s preferred feature space, i.e. identity or expression. However, different neurons have monotonic tuning curves with different slopes. This means that the distributed representation across a population of such neurons should still be sufficient to specify the exact identity or expression of a face stimulus. The reason why neurons were found to encode more information about identity than expression in our simulations might be related to the fact that with a set of realistic faces, such as the Ekman set (Friesen and Ekman, 1976), a pixel wise variation in identity tends to be greater than the variation in expression (Calder et al., 2001) so enabling easier discrimination for identity.

We next investigated what facial features, such as eyes, nose, and mouth, the different kinds of 4th layer neurons were responding to. This was done by tracing the connections that had been strengthened by learning from the 4th layer neurons back to the input Gabor input filters. Figure 32 shows the Gabor filters that have strong

connectivity through the network to example 4th layer neurons in VisNet which are individually tuned to one of four global attributes: Happy, Sad, Identity A, and Identity B. Each of the four corresponding subplots shows the Gabor filters with strong connectivity to that neuron as well as the neuron's firing rate responses to the facial stimuli representing combinations of identity and expression shown in Figure 29. It can be seen that the neuron tuned to Happy faces receives strong connectivity from Gabor filters representing the mouth. On the other hand, the neuron tuned to Sad faces receives strong connectivity from Gabor filters representing the eyes and eyebrows. Interestingly, the two neurons that differentiated between Identity A and Identity B received strong connectivity from Gabor filters representing the facial outline.

The above results indicate that there might be a relationship between neurons that represent particular global attributes, such as Happy, Sad, Identity A, and Identity B, and the kind of neurons discussed in the study 2 that encode the spatial relationships between local facial features such as the eyes, nose, and mouth with monotonic tuning curves. In fact, it could be that these apparently two different kinds of neuronal response characteristic is displayed by the same neurons. In other words, we wonder if the neuron that appears to respond to a global attribute such as Sad is simply responding to a particular spatial relationship between local facial features with a monotonic tuning curve.

To investigate this possibility, we took the four example cells shown in Figure 32, which represent particular global attributes such as Happy, Sad, Identity A, and Identity B, and applied the same analysis that was used in Study 2 for Figure 21 with the test faces shown in Figure 20. These results are shown in Figure 33, which shows how the firing rate responses of the four neurons vary with the spatial relationships between facial features. Each row shows the responses of a different neuron, while each column corresponds to a different kind of spatial relationship: inter-eye distance, eyebrow angle, eye height and mouth shape. The individual subplots show the firing rate responses of the neuron as the corresponding spatial relationship is varied across ten selected feature values. It can be seen that some neurons responding to global

attributes such as Sad and Identity A have monotonic tuning to particular spatial relationships between local facial features. For example, the cells that encode the facial expressions Happy and Sad essentially encode the shape of the mouth. While the cell tuned to Identity A encodes the eyebrow angle. The cell tuned to Identity B does not show a strong correlation to any of the four dimensions we tested, but it is quite possible that this cell encodes a different spatial relationship between facial features that is not shown here. These results strongly support the notion that a neuron that appears to respond to a global attribute is simply responding to a particular spatial relationship between local facial features with a monotonic tuning curve. Thus, these two kinds of neuron may in fact be the same, with neurons encoding different global attributes simply representing different spatial relationships between local features with monotonic tuning curves or particular combinations of them.

***Additional Study: the Representation of Six Basic Expressions.*** So far in this paper we have trained VisNet on only two different facial expressions, happy and sad, including intermediate expressions. This leaves open the question of whether VisNet could learn to recognise a larger number of different expressions. This question is in part motivated by a recent study by Sormaz et al. (2016), which has shown that the perceptual similarity of five expressions (happy, sad, angry, disgust, and fear) could be predicted from the patterns of neural response in the STS.

To address this question, we conducted an additional VisNet study where the network was trained on a set of 100 randomly generated facial identities for each one of six basic expressions: Happy, Sad, Anger, Disgust, Fear and Surprise. Then, the network was tested on a new set of randomly generated face stimuli for each of the 6 facial expressions. Specifically, for each expression, we created 10 different random facial identities in order to test whether the network representation of facial expression could generalise across the different facial identities. Figure 34 shows the results of the simulation. Each subplot in the top row shows the average responses of 10 cells, which are identified to carry the highest single cell information for a particular facial expression, to ten different randomly generated facial identities with that expression.

Although these results do not show perfect performance, it can be seen that the neurons shown in each subplot do respond more to faces with their preferred expression. The subplots in the middle row show the gabor filters that are most strongly connected to the ten output cells in the top row that represent each expression. It can be seen that the Happy neurons (first column) are receiving strong connections from gabor filters representing the shape of the mouth, while the Anger neurons (third column) are receiving strong connections from a different part of the mouth. The Sad neurons (second column) are receiving strong inputs from gabor filters representing the shape of the eyes, while Fear neurons (fifth column) and Surprise neurons (sixth column) receive strong inputs from different parts of the eyebrows.

**Learning mechanisms by which the network may form distinct representations of global facial attributes such as identity and expression : One layer network simulations.** The question is how the network can develop separate output representations of different global facial attributes such as identity and expression if these attributes are always seen together at the same time. Moreover, we wonder how the same retinal input neurons are used to encode the two global attributes simultaneously. Somehow, through a hierarchical series of neuronal layers, the primate visual system must use competitive learning to separate these global attributes, which are initially encoded by overlapping sets of retinal input neurons, onto distinct populations of output cells.

In order to explore the mechanisms by which this transformation might take place, we ran simulations of an idealised one-layer competitive neural network as described in Appendix A, but now with two 1-dimensional input spaces. One of the input spaces could be considered as encoding facial identity, while the other input space encoded facial expression. Each input space was represented by a 1-dimensional row of 100 neurons.

In some simulations the two input spaces were completely orthogonal to each other in that they shared no input neurons. In this case, there was a total of 200 input neurons. On the other hand in other simulations, the two input spaces shared some

neurons. In the case of completely overlapping input spaces the network contained a total of 100 input neurons, with these neurons ordered differently within the two spaces. The location of a face in each of these two spaces was encoded by the position of a Gaussian activity packet within that input space. The standard deviation  $\sigma$  governing the width of the Gaussian activity packets in both input layers was set to 10 in all simulations.

At each timestep during training, an input stimulus was defined by Gaussian activity packets presented at random locations within each of the two input layers. In simulations with dependent motion, the two input spaces were fully statistically linked in that the Gaussian activity packets always occurred at corresponding locations in the two spaces. In simulations with independent motion, the two input spaces were statistically independent in that the locations of the Gaussian packets in the two spaces were entirely independent of each other.

The activities of the input neurons were fed through the feedforward synaptic connections to drive the responses of 100 neurons in the output layer. Combined lateral inhibition and excitation was implemented between neurons in the output layer in order to effect competition. During training, the feedforward synaptic weights were then modified using a local associative (hebbian) learning rule with synaptic weight vector normalisation as described in Appendix A.

We explored how the output representations that developed in the network though learning were affected by (i) the degree of statistical independence between the two spaces during training, i.e. whether identity and expression varied independently of each other over the stimulus training set, and (ii) the degree of overlap between the input neurons encoding the two spaces, i.e. how many neurons the two input spaces had in common.

After training, we analysed the learned response behaviours of the output neurons using two methods. In the first method, the position of the Gaussian activity packet in one of the input spaces was systematically shifted through neurons 1 to 100, while the position of the Gaussian packet in the other input space remain fixed at the centre of



that space. In the second method, we presented Gaussian activity packets at all  $100 \times 100$  combinations of positions within the two input spaces, and the firing rate response table of each output cell was recorded for comparison with the results of the VisNet simulation reported in Figure 30.

Figure 35 shows how the *dependent* motion of Gaussian activity packets in two input layers during training affects the learned response properties of output neurons. In this simulation there was no overlap between the two input spaces. The top six subplots show the results of the first method of analysis. The three columns show the (a) weight matrix, (b) activation matrix, and (c) firing rate matrix of the population of output neurons. With dependent motion of the activity packets in the two input layers during training, the firing rate maps of the output neurons in response to the two input spaces largely overlap. Thus, the output neurons failed to develop separate representations of the two input spaces. The four subplots in the bottom row (d) show the second method of analysis. Each of the four subplots in the bottom row shows the firing rate responses for a different output neuron. Individual output neurons learned to respond to particular combinations of locations in the two input spaces that occurred together during training. Thus, these neurons had not learned to respond selectively to just one or other of the two input spaces.

Figure 36 shows how the *independent* motion of Gaussian activity packets in two input layers during training affects the learned response properties of output neurons. In this simulation there was again no overlap between the two input spaces. It can be seen in (a), (b) and (c) that the output neurons have developed separate representations of the two input spaces, with individual neurons responding to just one of the input spaces. In particular, the four output neurons shown in (d) each learned to respond selectively to a localised end region of one of the input spaces. For example, cell 82 shown in the first column of Figure 36(d) responds selectively to the right side of input space B regardless of where an activity pattern occurs in input space A. Similarly, cell 75 presented in the third column of Figure 36(d) responds selectively to the top of input space A regardless of the location of an activity pattern in input space B. Thus, the

output neurons successfully developed distinct representations of the two input spaces when the motion of the Gaussian patterns in the two input layers was independent.

Next, we tested the network by gradually increasing the overlap of the two input spaces. Figure 37 shows the results of a simulation with 50% overlap, while Figure 38 shows the results of increasing the overlap to 100%. In both of these simulations, the Gaussian activity patterns moved independently through the two input spaces during training. We found that even if the retinal input neurons are entirely overlapped, the output neurons still developed separate representations of two input spaces. This effect relied on the motions of the activity patterns in the two input spaces being independent during training.

We propose that these simulations may explain how the primate visual system develops physically separate representations of global facial attributes such as identity and expression, with individual neurons responding selectively to a localised region of one of these spaces, even though both attributes are encoded by the same population of retinal input neurons.

## Discussion

We have presented biologically plausible neural network simulations of the visually-guided development of facial representations in the visual brain using completely unsupervised learning mechanisms with feed-forward visual processing. These simulations contrast with many current engineering approaches based on the feedback of error signals from higher- to lower-levels of representation to guide supervised learning of facial attributes such as identity and expression (Lawrence et al., 1997; Lisetti and Rumelhart, 1998; Tsigman et al., 2014). Supervised learning by back-propagation of error (Rumelhart et al., 1986) is not a biologically plausible mechanism for learning facial representations in the brain. Although there exist back-projections in the visual system, it is not possible that these are carrying the kind of error signals needed by back-propagation of error learning (Stork, 1989). Hence, the simulations reported in this paper represent an important theoretical advance in

understanding how the visual system in the brain learns to represent the rich spatial structure of the faces.

In this paper, we conducted a series of simulation studies investigating how visual representations of faces may develop in the primate visual system. In particular, we trained an established hierarchical neural network model of the primate ventral visual stream, VisNet, with realistic human face stimuli constructed using FaceGen. As a result, we found that the network successfully developed various kinds of cells with response properties similar to those reported in neurophysiological studies. To further advance our understanding of the learning mechanisms involved, additional simulations were performed within simplified one-layer competitive network models.

Our initial simulations with the VisNet model showed the development of neurons that learned to respond to individual facial features such as the eyes and mouth, as well as combinations of these features, as has been reported in single cell recordings in the macaque brain (Freiwald et al., 2009). However, the question was how neurons might learn to respond to individual facial features if the facial features are always seen together within whole faces during training. Particular facial features such as the eyes occur in different shapes across different faces. Thus, across a population of faces the network will be exposed to different combinations of facial feature shapes on different occasions. This will lead to a statistical decoupling (Stringer et al., 2007; Stringer and Rolls, 2008) between the individual facial features, which we hypothesised may force the neurons in higher layers to learn to represent the individual features rather than whole faces. This hypothesis was confirmed in the VisNet simulations, where it was found that the output neurons switched to predominantly representing the individual facial features as the number of possible shapes of any facial feature  $p$  used to generate the set of training faces increased from 1 to 2.

We further hypothesised that as the number of shapes of any facial feature  $p$  increased further, an invariance learning mechanism known as continuous transformation (CT) learning would begin to drive the development of neurons that responded invariantly to many or all of the shape variations of a particular facial

feature. Such neurons would represent a facial feature such as a mouth irrespective of the particular shape of that feature. This hypothesis was also confirmed in VisNet simulations as  $p$  was increased to 5, 10 and 30. At  $p = 30$  there was a sharp rise in the number of neurons that responded to all 50 of the differently shaped eyes used to test the network.

Furthermore, the VisNet simulations also developed some cells with monotonically increasing or decreasing tuning responses to gradually changing spatial relations between facial features such as inter-eye distance, as has been observed in neurophysiology studies (Freiwald et al., 2009). The question was how such monotonic response properties develop. In complementary simulations of a one-layer competitive network, we found that the finite receptive field of a neuron due to a topologically restricted fan-in of afferent synaptic connections, as well as the nature of the competition within the output layer, both played important roles in the emergence of neurons with monotonic tuning.

#### **Relationships between the global facial representations and local facial feature representations**

We also found that VisNet developed neurons encoding global facial attributes such as face identity and facial expression as reported in neurophysiology studies (Morin et al., 2014). The question was how different sub-populations of higher layer neurons can learn to respond selectively to either face identity or expression if the network is always exposed to both attributes simultaneously, and the same retinal input neurons represent both global attributes simultaneously in a complex distributed manner. In complementary simulations of a one-layer competitive network, we showed that the network can develop separate representations of multiple perceptual input spaces such as facial identity and expression even if the input neurons encoding these spaces are fully overlapping. In particular, this may occur when the input patterns vary independently between the different input spaces. This result provides a possible mechanism for the simultaneous development of multiple global facial representations

1208 such as facial identity and expression.

1209         In the main simulation study reported in Study 3a, we showed that the cell that  
1210 learned to be selective to happy faces had a higher sensitivity to the shape of the  
1211 mouth. Interestingly, Gosselin and Schyns (2001) explored the specific visual  
1212 information humans use to recognize global attributes of faces based on a technique  
1213 called “bubbles,” and they also found that the humans use information around the  
1214 mouth for expression extraction. Their study has indicated that rather than the facial  
1215 features which simply have the highest local variance between the considered categories,  
1216 humans tend to use “partially efficient, not a formal, optimally efficient, feature  
1217 extraction algorithm” (Gosselin and Schyns, 2001). In the additional simulation study  
1218 conducted at the end of Study 3a, we have also presented that the Anger neurons (third  
1219 column) are also receiving strong connections from the mouth. The Sad neurons  
1220 (second column) are receiving strong inputs from gabor filters representing the shape of  
1221 the eyes, while Fear neurons (fifth column) and Surprise neurons (sixth column) receive  
1222 strong inputs from different parts of the eyebrows. These results would provide a  
1223 predictions about the facial features that might be used for the processing of facial  
1224 expressions in the brain.

1225         Furthermore, one of the most important arguments we raise is that the neurons  
1226 that encode global attributes of faces (such as facial identity and expression) and the  
1227 neurons that encode a spatial relationship between facial features (such as inter-eye  
1228 distance) are essentially the same. More specifically, we propose that neurons encoding  
1229 different global attributes such as expression simply represent different spatial  
1230 relationships between local features with monotonic tuning curves or particular  
1231 combinations of these spatial relations. In this way, the population response of a set of  
1232 facial features would be amplified for extreme compared with intermediate feature  
1233 values along the visual pathway, and thereby explain why faces with more deviant  
1234 appearances are recognized better than those which are more typical (Rhodes, 1997;  
1235 Benson and Perrett, 1991; Bruce and Young, 2011). In particular, this proposal  
1236 contrasts sharply with the idea of neurons being assigned in an entirely random

distributed manner to represent particular facial identities. Instead, neurons encoding facial identity are in fact representing specific structural information about the faces they encode. Our simulation results provide convincing evidence for this argument.

## The Representation of Faces and Non-face Objects

Recently, a modelling study carried out by Khaligh-Razavi and Kriegeskorte (2014) demonstrated that a number of unsupervised neural network models developed neuronal representations of faces that were highly correlated compared to the representations of non-face objects. These modelling results mirrored a similar effect found in actual data collected from monkey IT (Kriegeskorte et al., 2008b) and the human temporal lobe (Kiani et al., 2007). On the other hand, Khaligh-Razavi and Kriegeskorte (2014) showed that none of those unsupervised neural network models successfully captures the high correlations in the neuronal responses to non-face objects, which is also present in the brain.

In order to compare our results with those published by Khaligh-Razavi and Kriegeskorte (2014), we analysed activity within the network by computing *representational dissimilarity matrices* (RDM) (Kriegeskorte et al., 2008a) for each layer of VisNet. Figure 39 shows the RDMs computed in response to 50 faces and 50 non-faces for each layer of VisNet before training (left column) and after training (right column). These results show that, after training, the output (4th) layer of the network demonstrates neuronal activity patterns that are highly correlated in response to pairs of stimuli from within one of the stimulus categories, i.e. faces or non-face objects, but are decorrelated in response to stimuli from different categories. It can also be seen that this effect gradually increases through successive neuronal layers of the network.

This result contradicts VisNet’s poor performance reported in Khaligh-Razavi and Kriegeskorte (2014). However, this inconsistency can be explained by the way the network was trained and the size of the network simulated in their study. In Khaligh-Razavi and Kriegeskorte (2014), the model was trained with a trace learning rule over two stimulus categories: 442 ‘animated images’ (faces/bodies of

humans/non-humans) and 442 ‘inanimated images’ (natural/artificial objects). In theory, any visual stimulus in the same category should be associated together with the trace learning rule they implemented. This makes the network difficult to develop a representation that is exclusively dedicated to faces. Additionally, the size of the network they simulated was 16 times smaller than the network simulated in the current study, which may also have resulted in limiting the potential of the model. Accordingly, in contrast to the previously reported result in Khaligh-Razavi and Kriegeskorte (2014), our own simulations showed high correlations in the responses of the output layer of VisNet after training with objects from the same stimulus category whether faces or non-face objects. This implies a potential for models to develop similar kind of self-organisation to our brains through feedforward, unsupervised, visually-guided training.

Another important aspect of the visual processing can be found in the cortical structure of the brains. Even though the task of both face and object recognition is achieved along the ventral visual pathway, there is physiological evidence that faces and non-face objects are processed in distinct cortical areas. For example, it has been found that faces are preferentially processed in the occipital face area (OFA) (Pitcher et al., 2011) and several later cortical areas known as ‘face patches’ (Tsao et al., 2006). These effects were in fact seen in our simulations. When the network was trained on a mixture of faces and non-face objects, it was found that neurons that learned to represent faces tended to be clustered together within localised patches in the output layer, while neurons representing non-face objects were clustered within separate patches. This effect was due to the use of a self-organising map (SOM) architecture implemented within each layer of the model. In this case, the short range excitatory connections between neurons within each layer encouraged nearby neurons to learn to respond to similar stimuli. This was sufficient to lead to separate patches for faces and non-face objects.

Figure 40 shows maps of the selectivity of all 4th layer neurons to the faces and non-face objects before and after training. The selectivity measure was calculated for each cell as follows. First, we computed the average firing rate response  $\in (0, 1)$  of the

cell to all 150 faces and the average firing rate response to all non-face objects. If the average firing rate response to both categories of stimuli was less than 0.8 then the cell was deemed not responsive enough and the selectivity measure was set to zero. If the average firing rate response of the cell was greater than or equal to 0.8 for at least one of the stimulus categories then the selectivity measure was calculated by subtracting the average firing rate response to the non-face objects from the average firing rate response to the faces. Thus, the selectivity measure has a value near +1 (red) for a cell that is selective to faces and a value near -1 (blue) for a cell that is selective to non-face objects. Figure 40 shows that layer 4 developed large distinct patches of neurons that were selective for either faces or non-face objects after training. As mentioned above, this map like structure is reminiscent of the localised regions of face selectivity, called face patches, reported in neurophysiology studies (Gross et al., 1972).

Figure 40 shows some evidence of spatially structured selectivity to the two stimulus categories in the untrained network. However, this is simply due to neurons in different regions of layer 4 receiving different amounts of connectivity from different regions of the retina because of the topological feedforward connectivity through the layers. In this case, neurons in the centre of layer 4 may be driven more by the relatively rich visual structure at the centre of non-face objects, while the neurons surrounding the centre of layer 4 may be driven more by peripheral facial features such as the facial outline, etc. This effect is also artifactual because it only arises due to the face and non-face objects not being shown in different retinal locations. However, the strengths of the feedforward connections are extensively modified during visually guided training, leading to the development of selectivity maps with relatively large, contiguous patches of stimulus selectivity, as seen in the primate visual system.

Even though faces and non-face objects are processed in different visual areas, does this mean that the underlying nature of visual processing is different for these two stimulus categories? It was originally suggested by Biederman and Kalocsai (1997) that the processing of faces is unlike that of other objects. These authors argued that the retinal image of an object is decomposed into simple 3D primitives called geons as well



as the spatial relationships between these primitives (Biederman, 1987). Such a structural description is viewpoint-independent. On the other hand, the information required for face recognition is proposed to be more holistic, which may be coded as a form of graph with each node representing a particular facial feature and each link representing a relationship between features (Lades et al., 1993; Biederman and Kalocsai, 1997).

However, it has later been argued that the basic visual processing of faces and non-face objects may in fact be similar (Mangini and Biederman, 2004; Yue et al., 2006). In particular, the different psychophysical behaviour towards faces and non-face objects may naturally arise if the representation of faces retains aspects of the original spatial filter representation. In particular, they proposed that larger receptive fields which partially overlap with each other would provide sufficient information to produce the sensitivity to the layout and the spacing of nameable parts of the face (configural effects). A recent simulation study conducted by Xu et al. (2014) supports this hypothesis and concluded that “the configural effect is largely a function of the overlap in the encoding of multiple face features allowed with large receptive fields”.

The simulations reported in this paper use the VisNet architecture originally developed by Wallis and Rolls (1997). This model uses a biologically plausible architecture, with unsupervised learning mediated by local, associative learning rules. Wallis and Rolls (1997) proposed that the learning principles underlying the processing of faces and non-face objects are similar, and used VisNet to model the development of transform invariant representations of both faces and non-face objects. These authors hypothesised that it would be possible for the model to develop detailed representations of the local features of faces if more neurons were incorporated into the model. We have now verified this prediction in a network with 16 times as many neurons as that originally used by Wallis and Rolls (1997).

## Limitations and Future Directions

The simulations reported in this paper used face images constructed using the FaceGen 3D face modelling software. In future work, we plan to replicate these studies using real faces. This will introduce new problems such as how the network achieves correspondence of the same facial features, such as the eyes or nose, across very differently shaped faces. We anticipate that this may require training VisNet on faces in different retinal locations and different scales to achieve representations of facial features that are both location and scale invariant. This will be a significantly more challenging task than with the controlled artificial FaceGen images used in the current study.

The version of the VisNet architecture used in this paper incorporated only bottom-up (feedforward) connections between successive layers of the network. No top-down connections were included in the model even though these are known to exist in the primate ventral visual pathway and are proposed to have a role in matching incoming inputs with top-down expectations or predictions (Clark, 2013). Nevertheless, the rationale for using this simplified architecture in the current study was that it is sufficient to replicate how neurons in face areas are able to learn to encode the various kinds of face related information.

However, Zhou et al. (2000) have shown that the responses of neurons in earlier stages of visual processing such as V1 and V2, which have preferred responses to oriented edges, are also modulated by which side of a figure the edge occurs on. This is the case even when the figure/background cues lie well outside the classical receptive field of the neuron. This suggests that global image context specifying border ownership modulates the activity of these neurons. This contextual information must be conveyed to these early stage visual neurons by some combination of top-down connections between layers and recurrent connections within layers. These observations imply that top-down connections do play a role in modulating the responses of neurons in earlier layers. However, it remains to be seen what role they may play in face processing. Indeed, we emphasise again that all of the representations that developed in the VisNet simulations described in this paper, which reflect the neuronal firing properties observed

in neurophysiology studies such as Freiwald et al. (2009), self-organised using purely feedforward processing.

Another thing to be mentioned is that the current study proposes theories that may explain the development of various neuronal properties that are localized along the ventral visual system, which may be mapped onto the ‘core system’ in the model proposed by Haxby et al. (2000); however, our model does not explicitly model cognitive processes, which is achieved in the ‘extended system’ to act in concert with the regions of the core system to extract meaning from faces (Haxby et al., 2000). Therefore, even though our results are compared with physiological data, such as face feature space representations (Freiwald et al., 2009) and global representations of identity and expression (Hasselmo et al., 1989; Morin et al., 2014), the model does not generate the behaviours associated with the extracted information.

We believe that such behaviour can be implemented with architectural extensions to the current model that may more accurately reflect the known neuroanatomy of the relevant brain areas. For example, Rolls and Treves (1998) have previously hypothesised that pattern association learning may operate in the feedforward connections from area TE at the end of the ventral visual pathway, which represents faces and other visual objects, to areas such as the amygdala and orbitofrontal cortex (OFC). Consistent with this theory, single unit recording studies have shown that neurons in the amygdala and OFC learn associations between visual stimuli (conditioned stimuli) and the corresponding tastes (unconditioned stimuli). Therefore, pattern association appears to operate in these brain areas, which are thought to be involved in the evaluation of the emotional valence of visual stimuli. This would be a useful extension of the model in order to compare the simulated results with human performance on a human categorization task.

Nevertheless, as Wallis (2013) explains, the work presented in this paper also “serves to explain how such a core system would operate, in terms of its adaptive encoding of objects of expertise, but not how these other systems come to extract information from it to solve specific tasks.” For example, Yankouskaya et al. (2014) has

recently reported that the level of integration of identity and emotion cues in faces may be determined by life experience and exposure to individuals of different ethnicities. This is consistent with the finding reported in Wallis (2013) that showed that the network trained on Caucasian faces exhibits less sensitivity to changes in appearance of Japanese faces than those of Caucasian faces. We have also shown that after the exposure to 450 faces with randomly generated identities and expressions, many cells became sensitive to changes of identity and expression in a target face. Moreover, the fact that some cells in our model became sensitive to both of these attributes is consistent with the physiological evidence provided by Morin et al. (2014). Such neuronal representations developed in the self-organizing models provide important information to the ‘external system’ to generate the perceptions and behaviour reported in cognitive experiments (Yankouskaya et al., 2014). Accordingly, this paper investigated the developmental process of various kinds of such ‘structural codes’ (Bruce and Young, 1986), which may set the necessary foundation to achieve face perception in the later stages.

**Acknowledgment**

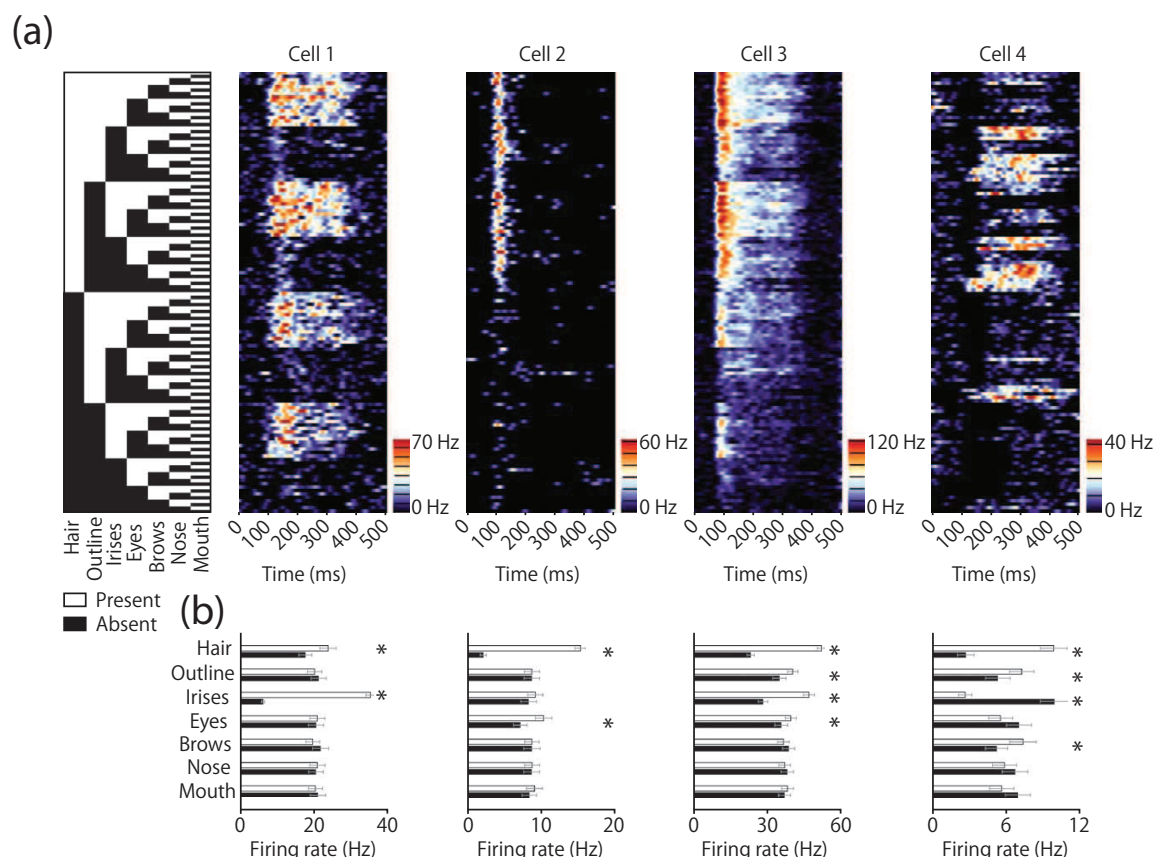
1422

1423       The authors wish to thank B.M.W. Mender and B.D. Evans for invaluable  
1424 assistance and discussion related to the research.

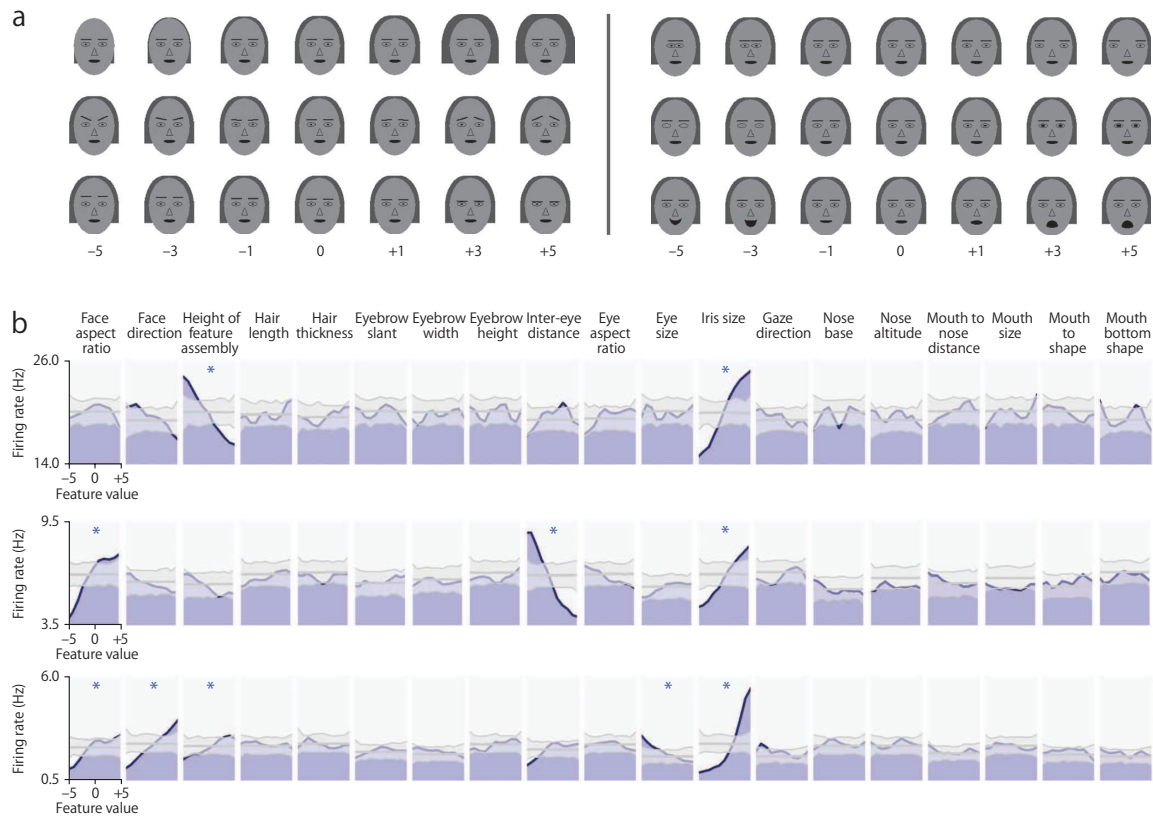
Table 1

*Parameters used for simulations*

Parameter	Value			
(a) VisNet				
Gabor: Phase shift ( $\psi$ )	0, $\pi$			
Gabor: Wavelength( $\lambda$ )	2			
Gabor: Orientation( $\theta$ )	0, $\pi/4$ , $\pi/2$ , $3\pi/4$			
Gabor: Spatial bandwidth ( $b$ )	1.5 octaves			
Gabor: Aspect ratio ( $\gamma$ )	0.5			
No. of Layers	4			
Retina	$256 \times 256 \times 16$			
	1st layer	2nd layer	3rd layer	4th layer
Dimension	$128 \times 128$	$128 \times 128$	$128 \times 128$	$128 \times 128$
Num. of fan-in connections	201	100	100	100
Fan-in radius	8	8	12	16
Sparseness of activations	2 %	20 %	30 %	30 %
Sigmoid slope ( $\beta$ )	190	40	75	26
Learning rate ( $k$ )	1.0	1.0	1.0	1.0
Training Epochs	50	100	100	76
Excitatory Radius ( $\sigma_E$ )	1.4	1.1	0.8	1.2
Excitatory Contrast ( $\delta_E$ )	5.35	33.15	117.57	120.12
Inhibitory Radius ( $\sigma_I$ )	2.76	5.4	8.0	12.0
Inhibitory Contrast ( $\delta_I$ )	1.6	1.5	1.6	1.5
(b) Simplified Network				
No. of cells in each layer	100			
Sigmoid slope $\beta$	10			
Learning rate $k$	0.001			
Training epochs	3000			
Sparseness of activations	50 % (simulation 1) and 25 % (simulation 2)			
Inhibitory contrast $\delta_I$	0.01			
Inhibitory radius $\sigma_I$	15			
Excitatory contrast $\delta_E$	0.5			
Excitatory radius $\sigma_E$	5			

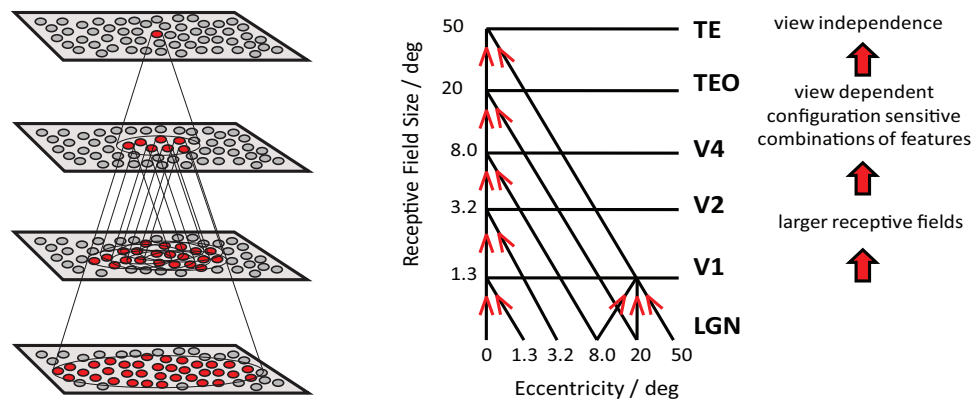


*Figure 1.* Physiological evidence from a single unit recording study carried out by Freiwald et al. (2009) showing neuronal selectivity for face parts in the primate ventral visual pathway. In this study, cartoon faces were shown to a macaque while the responses of neurons in the middle face patch were recorded. The face stimuli were varied across trials by varying which combination of facial features was present. The top panel (a) shows which facial features were present on each trial (left), and the corresponding responses of four example cells. All combinations of seven face parts (hair, outline, irises, eyes, eyebrows, nose and mouth) were shown, including the whole cartoon face with all features (top row) and a gray background without any face features (bottom row). The responses of the four example cells to each of the face stimuli are shown as a function of time. The bottom panel (b) shows the average neuronal responses in the presence (white bars) or absence (black bars) of a given face part. \* indicates significant modulation. Cell 1 fired significantly more strongly when irises were present and when hair was present. Cell 2 was influenced by two facial features, and cells 3 and 4 by four facial features. Cell 4 responded more strongly when irises were absent than when they were present. In cell 4, interactions between face parts were stronger than in the other cells, giving rise to less regular responses across stimulus conditions. Figures are excerpted with permission from Freiwald et al. (2009).

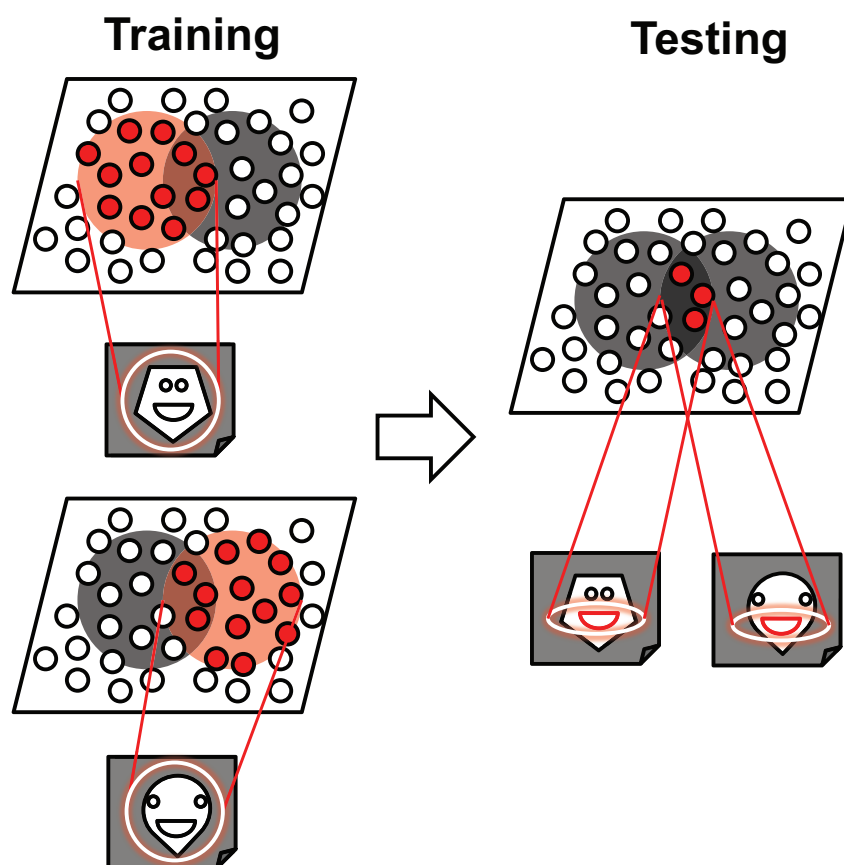


*Figure 2.* Physiological evidence from the single unit recording study carried out by Freiwald et al. (2009) showing neuronal selectivity for the spatial relationships between facial features with monotonic tuning curves. The top panel (a) shows example cartoon face stimuli for six different feature dimensions (hair width, eyebrow slant, eyebrow height, inter-eye distance, iris size and mouth shape) with seven feature values each spanning the entire range of values. The bottom panel (b) shows the response curves of three example cells to each of 19 feature dimensions. For each of the feature dimensions, the response curve (blue) is shown at a delay corresponding to maximal modulation. Maximal, minimal and mean values from the shift predictor are shown in gray. Asterisks mark significant modulation. Figures are excerpted with permission from Freiwald et al. (2009).

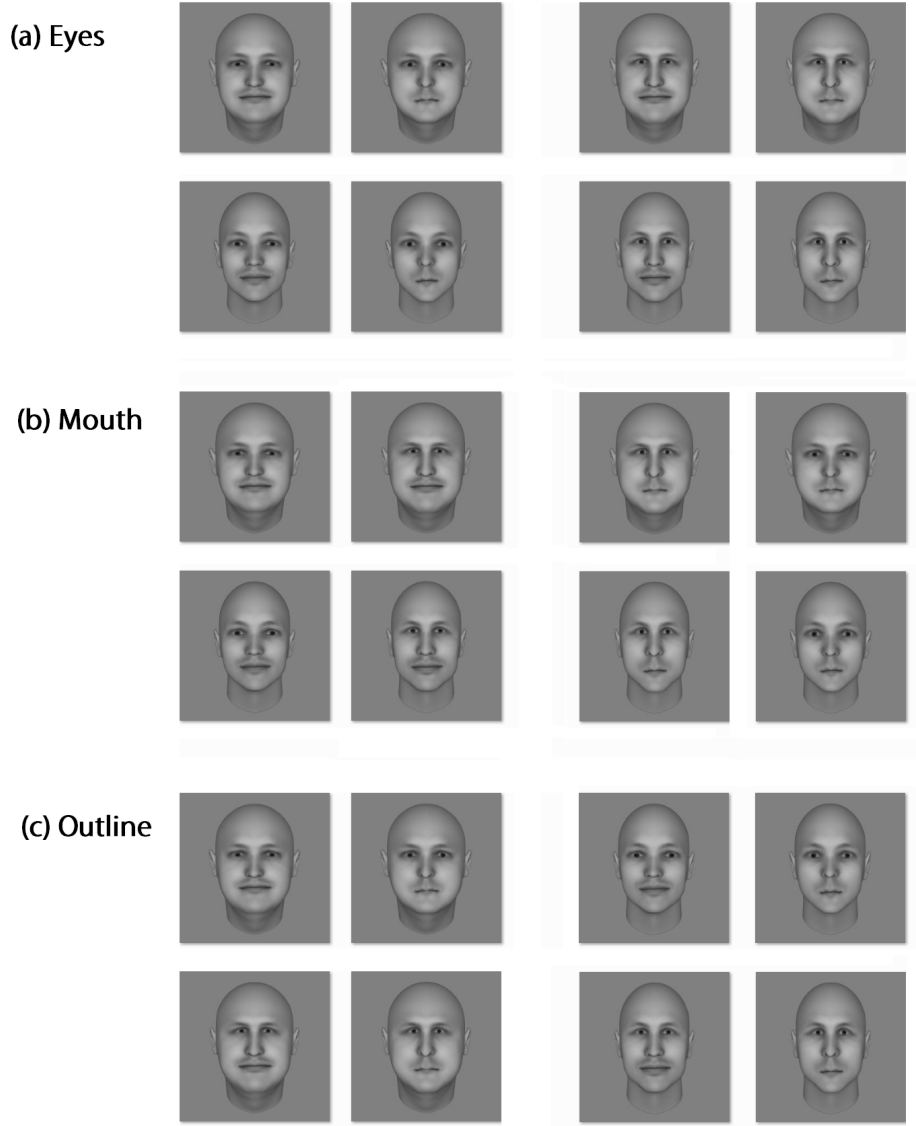




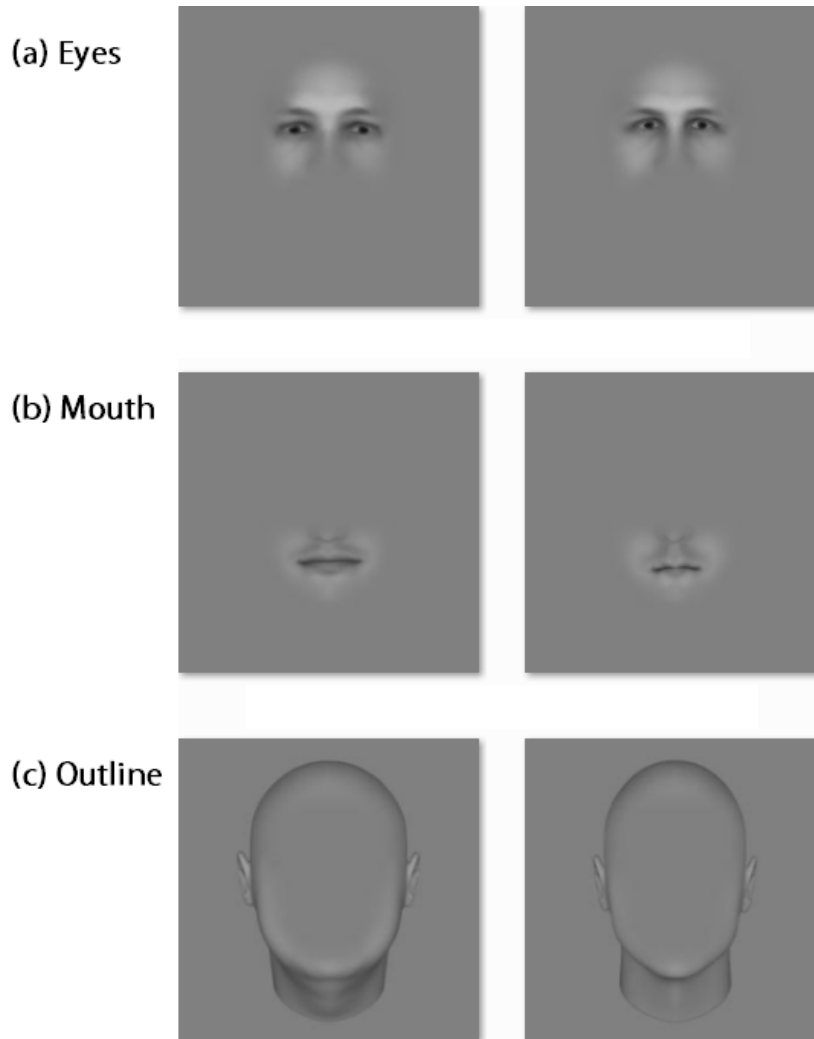
*Figure 3.* Left: Stylised image of the four layer VisNet architecture. Convergence through the network is designed to provide fourth layer neurons with information from across the entire input retina. Right: Convergence in the visual system V1: visual cortex area V1; TEO posterior IT, TE anterior IT



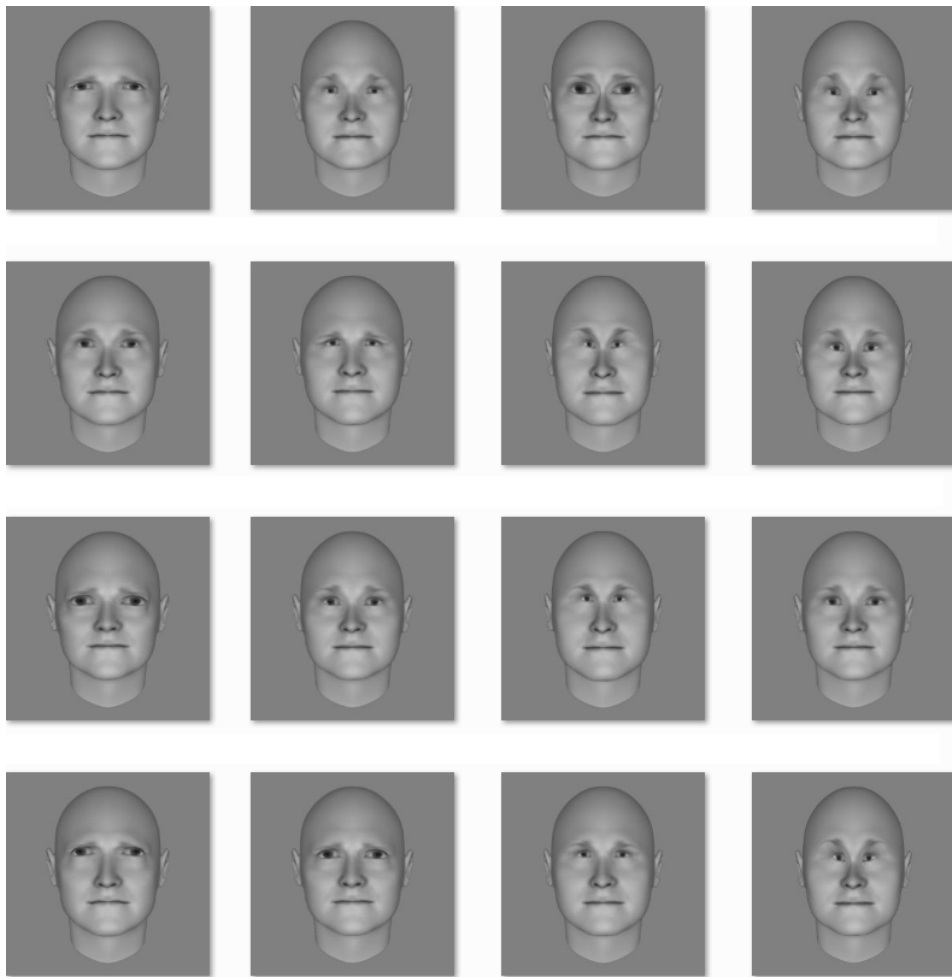
*Figure 4.* Illustration of how the network model develops neurons that have learned to respond to individual facial feature (mouth) from whole faces via statistical decoupling, which is similar to the mechanism of which V4 neurons may learn to represent the shapes of local boundary elements (Eguchi et al., 2015). Left: during training, the network is presented with many different faces, where each shape is defined by a unique combination of facial features of different shapes. Two such faces are shown here. Each of these faces stimulates a different subset of neurons in the output layer of the network. The two faces shown have the mouth in common. As a result, this mouth becomes especially strongly connected, through associative learning in the feed-forward synaptic connections, with the intersection of the two subsets of output neurons shown. This intersecting subset of neurons will come to represent the mouth of the particular shape in the two faces. Right: during testing, whenever the mouth is part of a face, the same intersecting subset of output neurons will be activated. A similar learning process will drive the development of many other subsets of output neurons representing different individual facial features.



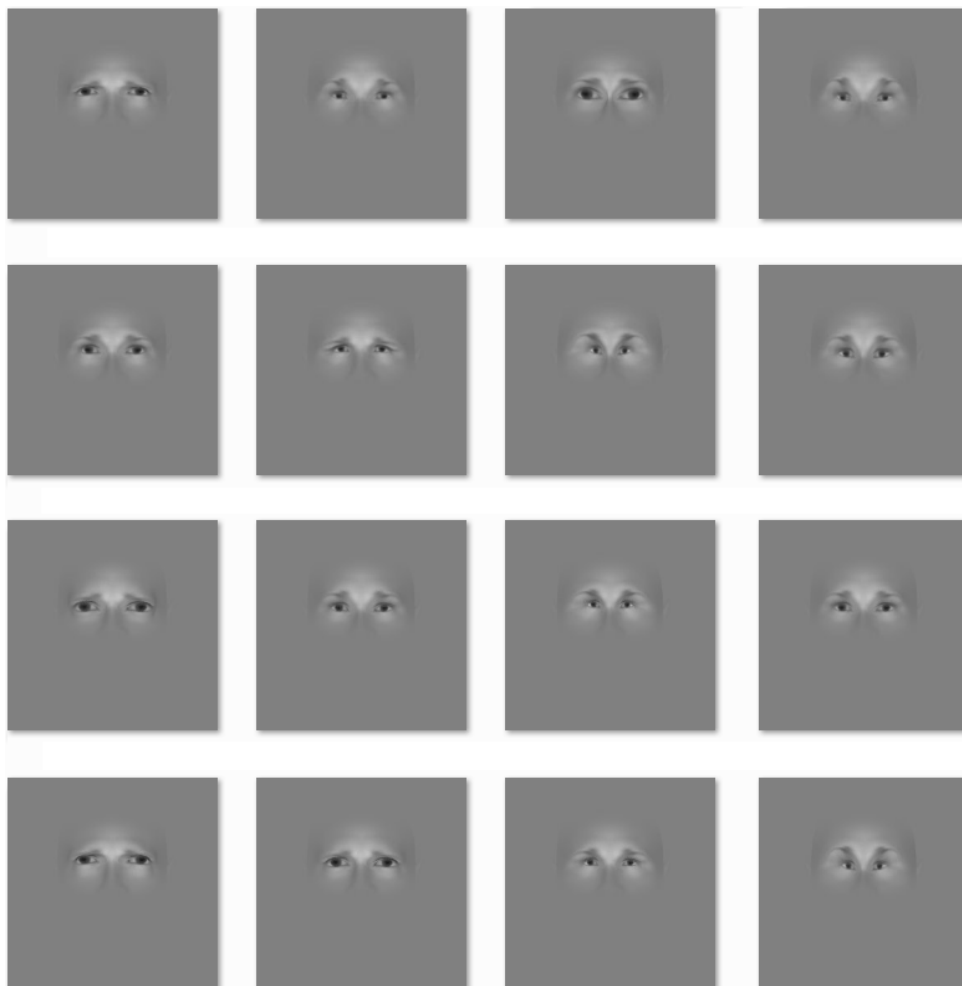
*Figure 5.* Example set of realistic faces used to train VisNet. This set of faces was generated by systematically varying the  $n = 3$  facial features eyes, mouth, and facial outline. In this example, there are  $p = 2$  shape variations of each facial feature. This gives a total number of  $p^n = 8$  faces that may be constructed by combining the facial features in different combinations. The top two rows show how the shape of the eyes is varied. These two rows show all possible  $p^n = 8$  faces, where faces with the first shape of the eyes are shown on the left and the faces with the second shape of the eyes are shown on the right. Similarly, the middle two rows show how the shape of the mouth is varied, where faces with the first mouth shape are shown on the left and faces with the second mouth shape are shown on the right. Lastly, the bottom two rows show how the facial outline is similarly varied between two different shapes shown on the left and right.



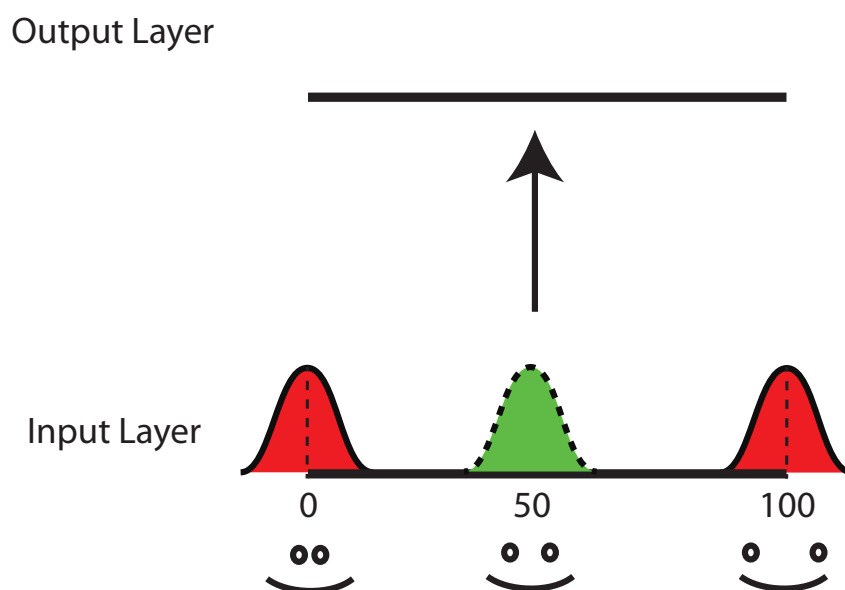
*Figure 6.* Example set of stimuli used to test VisNet after training. Each of the test stimuli is constructed by extracting one of the facial features used during training. In this example, each test stimulus contains one of the  $n = 3$  facial features: (a) eyes, (b) mouth, and (c) facial outline. Each of the four facial features has  $p = 2$  possible shape variations.



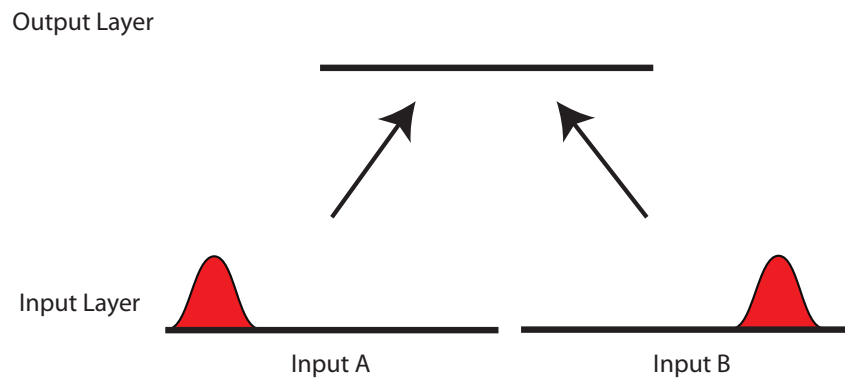
*Figure 7.* Example set of realistic faces used to train VisNet. This set of faces was generated by varying the shape of the eyes.



*Figure 8.* Example set of faces used to test VisNet. This set of faces only contained eyes, and was generated by varying the shape of the eyes between faces.

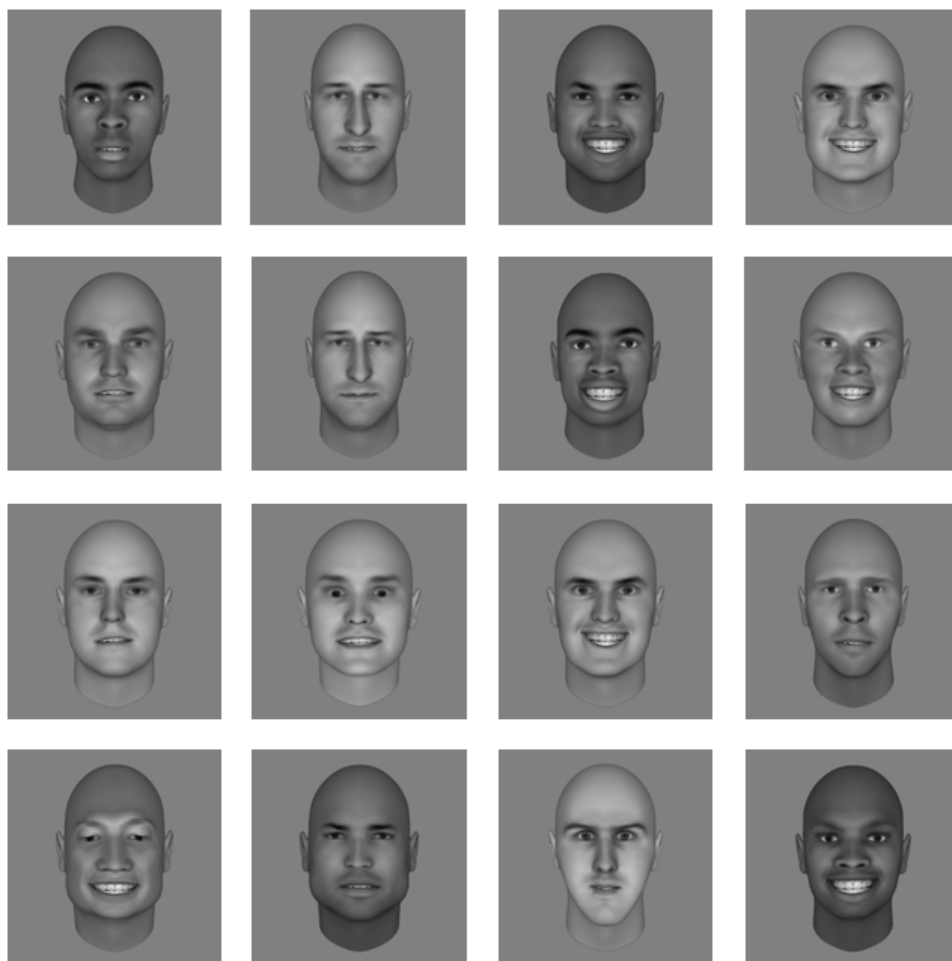


*Figure 9.* Figure showing the architecture of simple one layer network where Gaussian packet of neural activities in the input layer represent a finite 1-dimensional feature space such as the distance between the eyes. The curve filled with green shows the situation where a whole input packet is within the input layer whereas the curves filled with red show the situations where only part of the input packet is within the input layer.

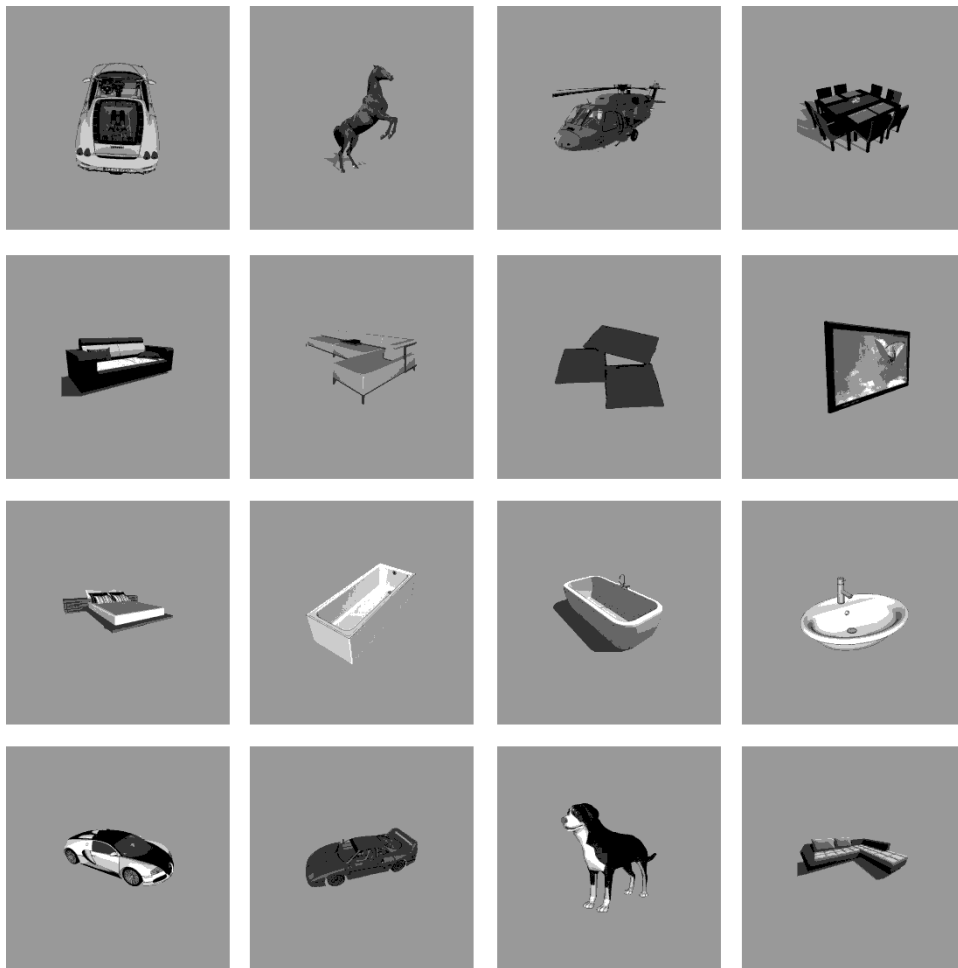


*Figure 10.* The architecture of a competitive neural network where the output neurons receive connections from two distinct populations of input neurons, A and B, which represent two different feature spaces such as facial identity and expression. The degree of overlap between the two input populations was varied across different simulations: 0% overlap, 50 % overlap, and 100 % overlap.

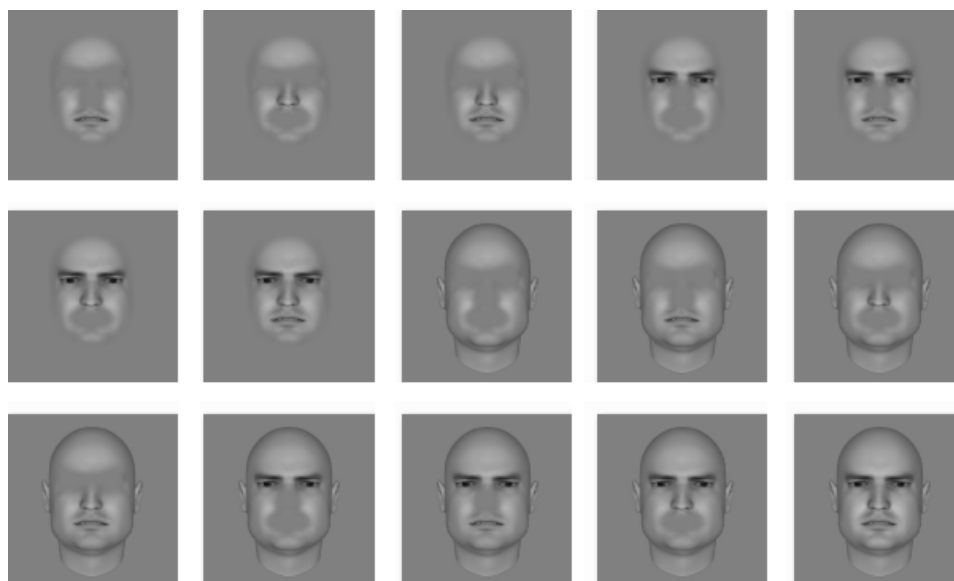




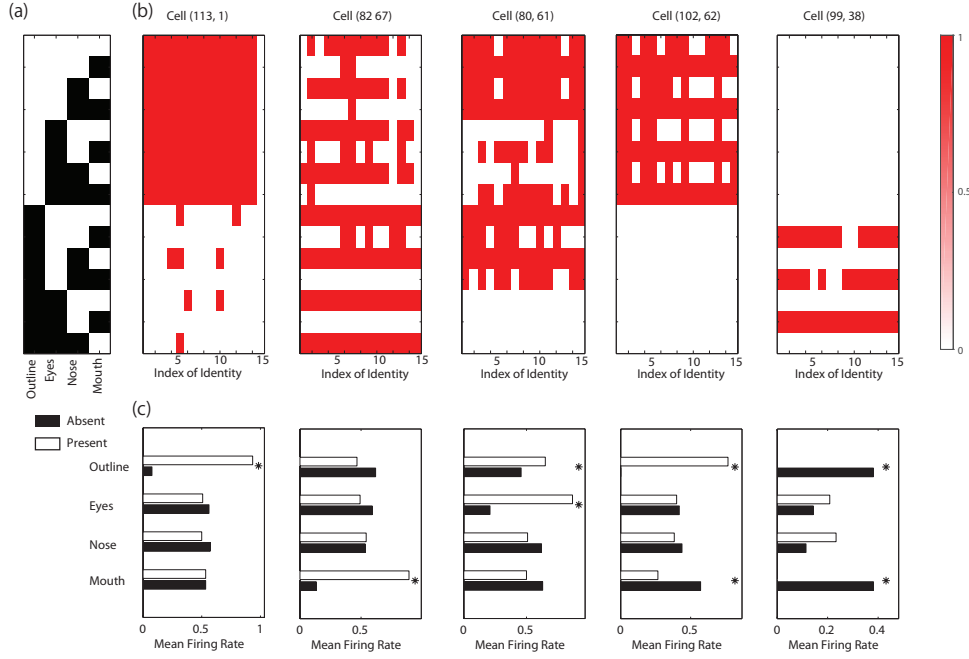
*Figure 11.* Examples of randomly generated faces with different identities used for training the VisNet network. These stimuli were generated using the commercial software FaceGen. The facial expression of each face was also randomly set along a continuous dimension between happy and sad.



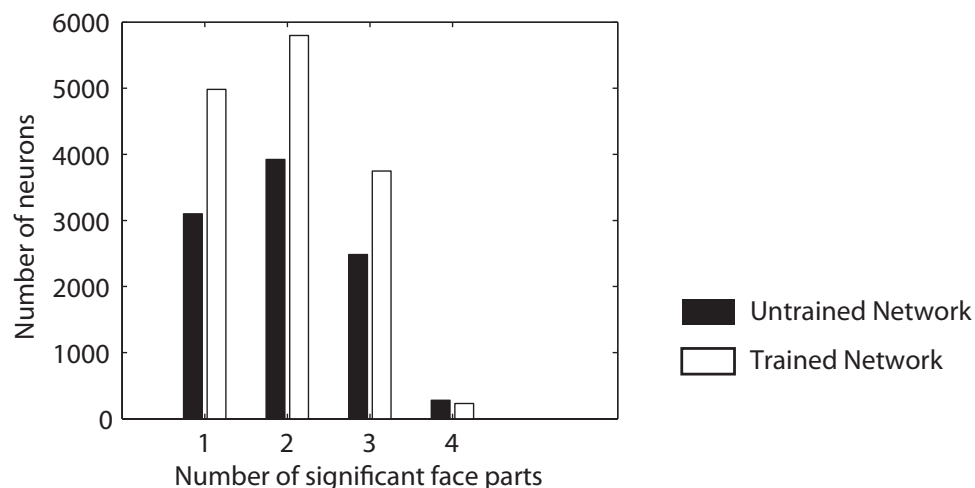
*Figure 12.* Examples of non-face objects used for training the VisNet network. Original images were retrieved from google 3D warehouse and then rescaled and grayscaled.



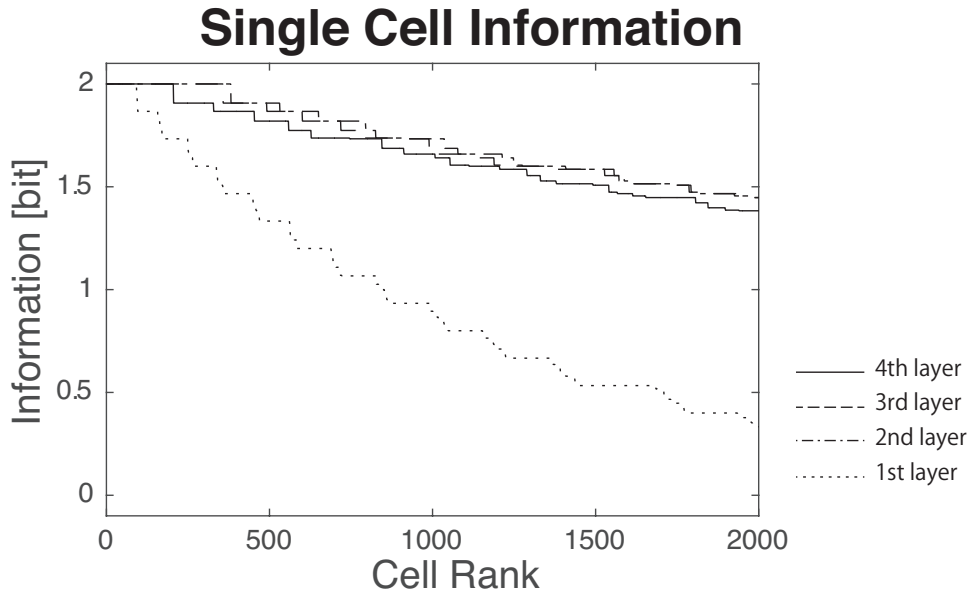
*Figure 13.* Examples of test faces which are composed of a different combination of the four facial features: facial outline, eyes, nose, and mouth. Different faces may have either 1, 2, 3, or 4 of these features present. The purpose of these face stimuli is to test for the existence of neurons that have learned to respond selectively to particular subsets of these facial features.



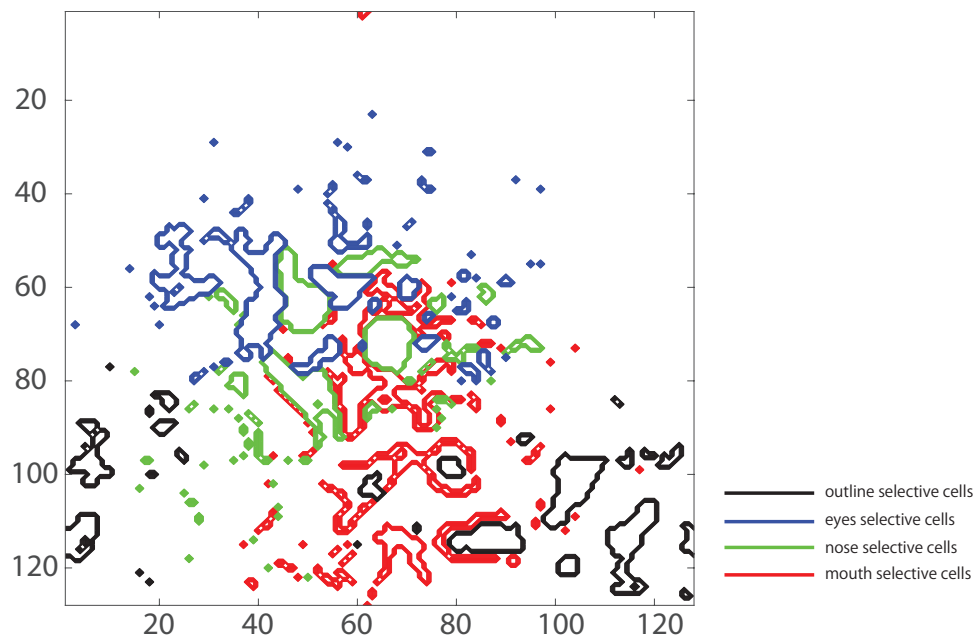
*Figure 14.* Simulation results showing the presence of 4th layer neurons that have learned to respond selectively to particular combinations of the four facial features: facial outline, eyes, nose, and mouth. The network was tested with faces constructed from all possible combinations of the four facial features. (a) The top left subplot shows the different combinations of facial features used to test the network, where each row corresponds to a different combination of facial features. (b) The five subplots on the top right show the responses of five different 4th layer neurons to faces constructed from different combinations of features. Each row corresponds to a different combination of facial features defined by the top left plot, and each column corresponds to a different facial identity. (c) The five subplots on the bottom show the average responses of the same 4th layer neurons to face stimuli with a given facial feature (white bars) and without the facial feature (black bars). Based on paired t-test, \* indicates significant excitatory modulation of the neuronal responses by a particular facial feature ( $P < 0.005$ ). For example, cell (113, 1) fired significantly more strongly when the facial outline was present ( $P < 0.005$ ).



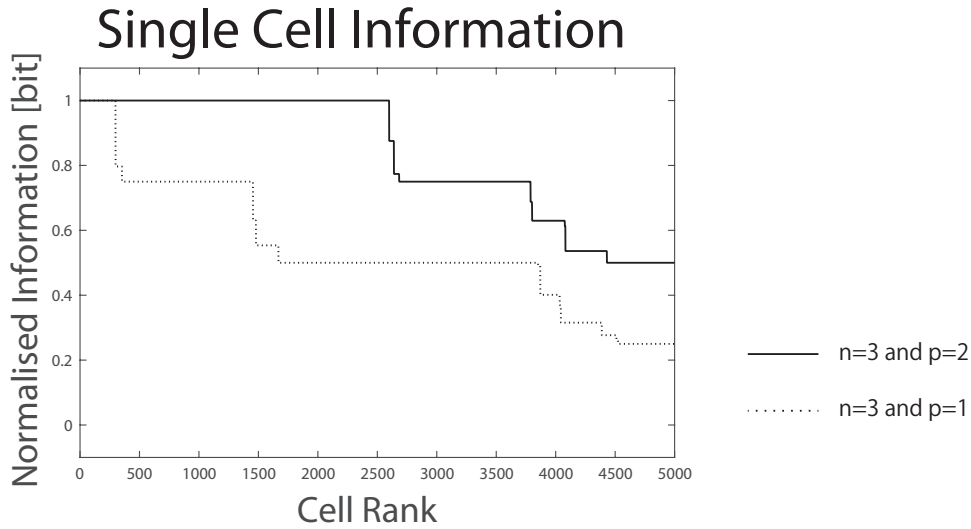
*Figure 15.* Frequency histogram showing the number of 4th layer neurons that responded significantly more strongly ( $P < 0.005$ ) to the presence rather than absence of a particular number (1, 2, 3, or 4) of facial features. The network was tested as follows: For each of the four facial features (outline, eyes, nose and mouth), we recorded the average response of each 4th layer neuron to (i) all possible faces that contain the feature and (ii) all possible faces that omit that feature. Then we computed for each neuron whether it responded significantly more to the presence rather than absence of that feature. This was done using a paired t-test. We repeated this procedure for all four facial features. Then, for each neuron, we recorded the number of facial features for which the neuron responded significantly more to the presence rather than absence of that feature. The histogram shows the number of neurons that responded more strongly to either 1, 2, 3 or 4 facial features. Results are shown before and after training.



*Figure 16.* Information analysis of selectivity of neurons to specific face parts. VisNet was trained on 450 realistic human faces as shown in Figure 11 and 150 non-face objects as shown in Figure 12 and then tested on four different face parts (mouth, nose, eyes, and outline) for 15 distinct facial identities shown in Figure 13. In this analysis, we tested whether individual cells had learned to respond selectively to the presence of a particular face part. To do this, we measured the amount of single cell information carried by cells about whether one of the four face parts was present in the test image. The figure shows single cell information plots for different layers of VisNet: 1st layer (dotted line), 2nd layer (dash-dot line), 3rd layer (dashed line), and 4th layer (solid line). Since there are four different face parts, the maximum amount of information possible is  $\log_2(4)$ , that is 2 bits. The results show that the number of cells that reached maximum single cell information is small in the first layer, is largest in the second and third layers, and then declines slightly in the fourth layer.

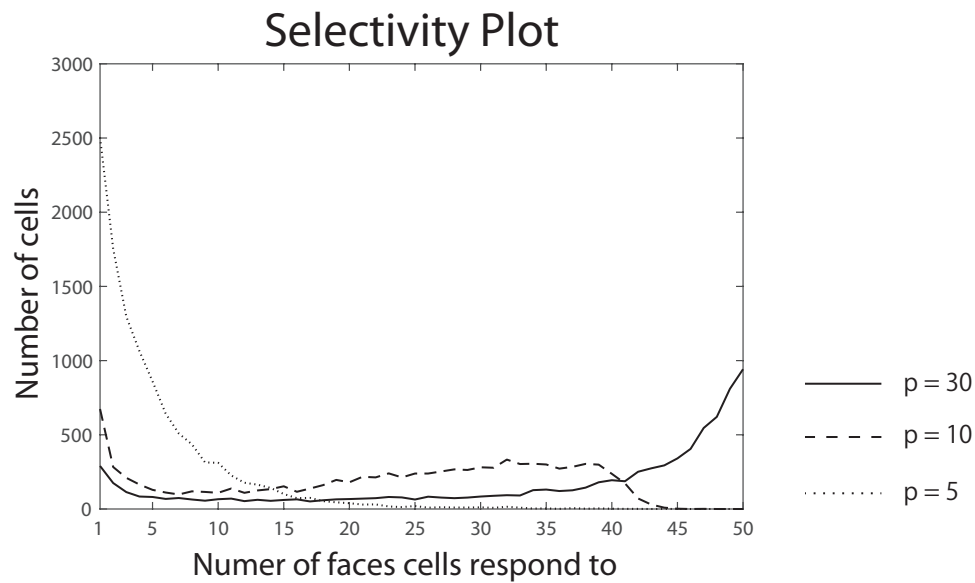


*Figure 17.* Map showing face feature selectivity of all 4th layer neurons to the mouth (red), nose (green), eyes (blue), and outline (black) features shown in Figure 13. The selectivity measure was computed based on single cell information analysis, and the 500 cells that carry the highest information for each facial feature are presented.

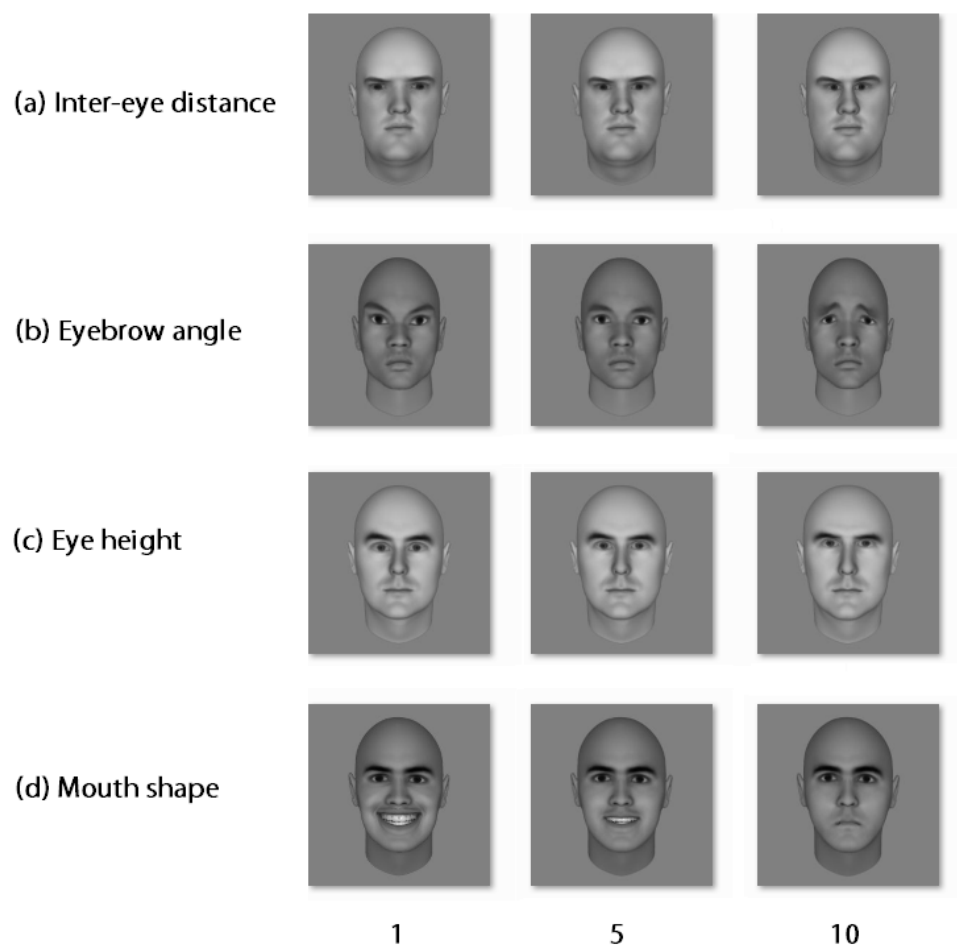


*Figure 18.* Results of simulations in which VisNet was trained on facial stimuli that were constructed by varying  $p$ , the number of possible shapes of each of the facial features. The face stimuli had  $n = 3$  facial features, eyes, mouth, and outline, each of which was varied over  $p$  different shapes during training. (An example set of training stimuli where  $p = 2$  is shown in Figure 5.) There were two separate simulations in which the training stimulus set had  $p$  equal to either 1 (dotted line) or 2 (solid line). In both simulations, the network was tested on the set of stimuli constructed by extracting only one facial feature from the facial stimuli used during training, as shown in Figure 6 for  $p = 2$ . The figure shows single cell information plots for the two separate simulations with  $p$  equal to 1 or 2. Each plot shows the normalized single cell information carried by each of the 3rd layer neurons (in rank order) about the presence of one of the  $n \times p$  facial features. The maximum amount of information possible for the two simulations is  $\log_2(n \times p)$ , that is 1.6 and 2.6 bits for simulations with  $p$  equal to 1 or 2 respectively. The results show that the number of cells that reached maximum single cell information increases as  $p$  increases from 1 to 2. This supports the hypothesis that the increased statistical decoupling between facial features across multiple faces as  $p$  increases from 1 to 2 forces neurons to learn to become more selective to particular facial features.

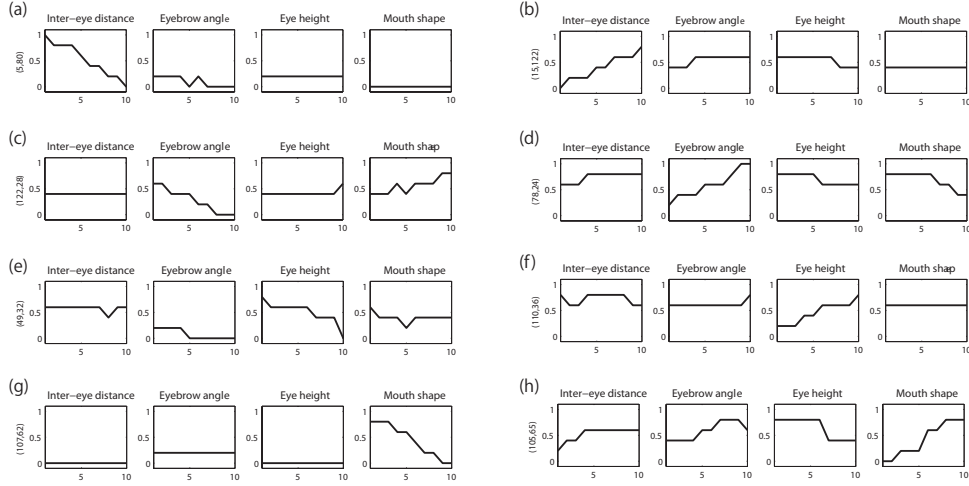




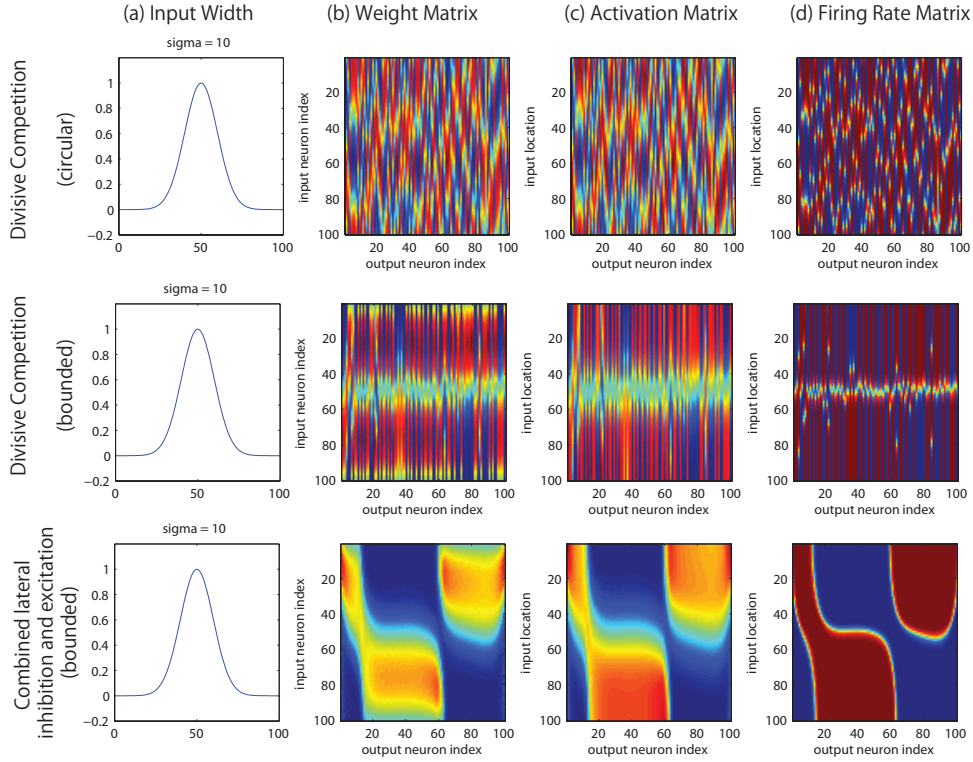
*Figure 19.* Results of simulations in which VisNet was trained on facial stimuli that were constructed by varying the number of possible shapes of the eyes as shown in Figure 7, and tested on facial feature stimuli that were constructed by extracting just the eyes of novel faces as shown in Figure 8. The plot shows how, as the network is exposed to more shapes of eyes during training, many cells start to exhibit shape invariant selectivity to the eyes.



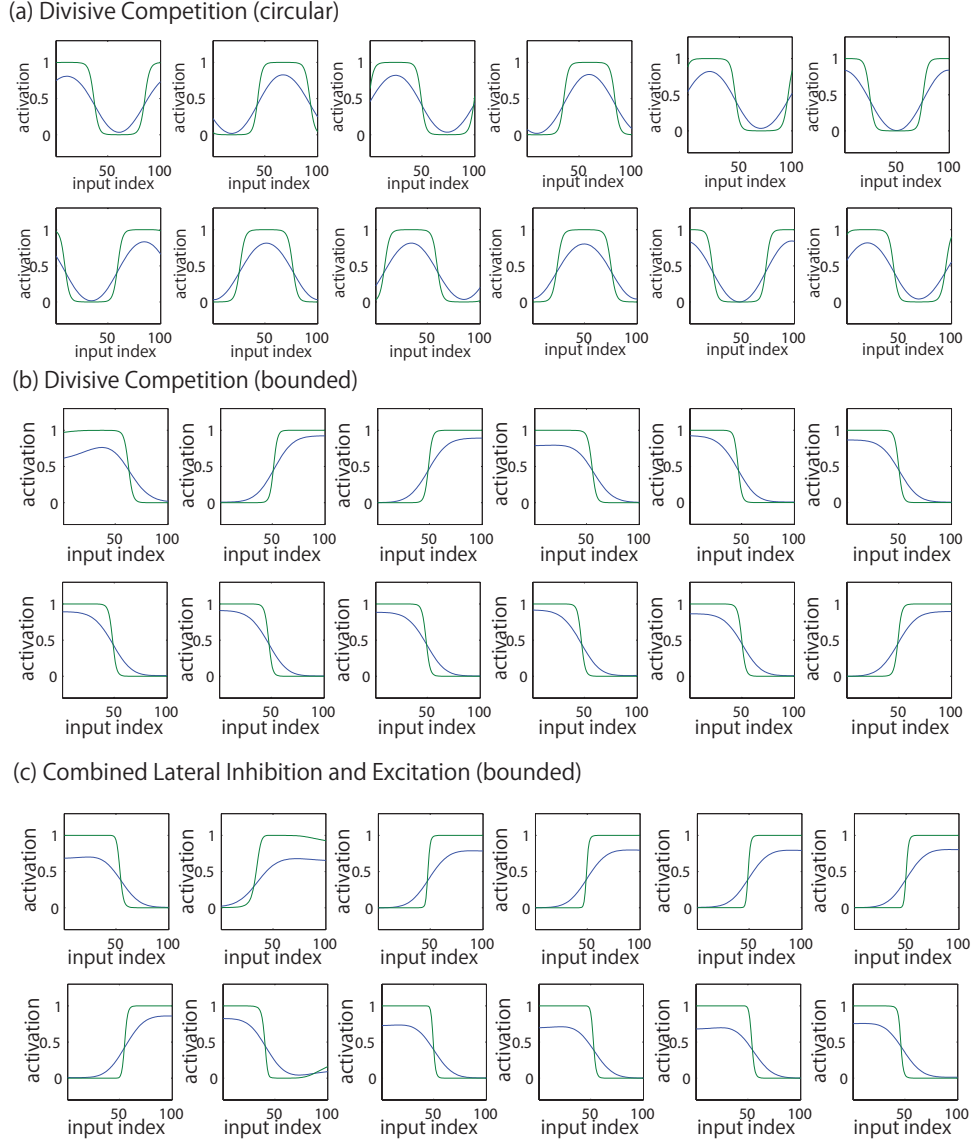
*Figure 20.* Examples of face stimuli used to test VisNet for the presence of neurons that had learned to represent the spatial relationships between facial features with monotonic tuning curves. Four different spatial relationships between facial features were varied during testing: inter-eye distance, eye-brow angle, eye-height and mouth shape.



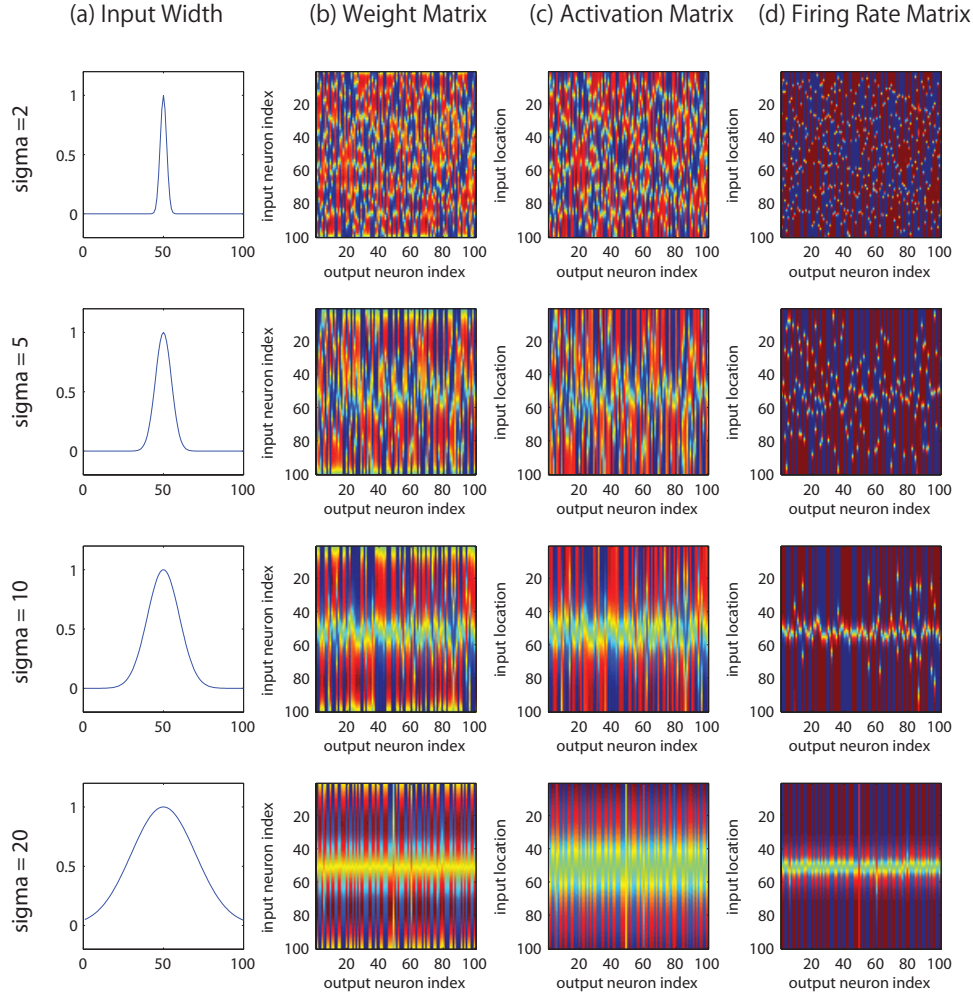
*Figure 21.* Firing rate responses of eight example neurons found in the 4th layer of VisNet which represent different spatial relationships between facial features with monotonic tuning profiles. The network was tested on the face stimuli shown in Figure 20. Each row shows the responses of a different neuron (a-h), while each column corresponds to a different kind of spatial relationship: inter-eye distance, eyebrow angle, eye height and mouth shape. The individual subplots show the firing rate responses of the neuron as the corresponding spatial relationship is varied across ten selected feature values. It is evident that the cells are tuned monotonically to different spatial relationships between the facial features: (a,b) inter-eye distance, (c,d) eyebrow angle, (e,f) eye height, and (g,h) mouth shape.



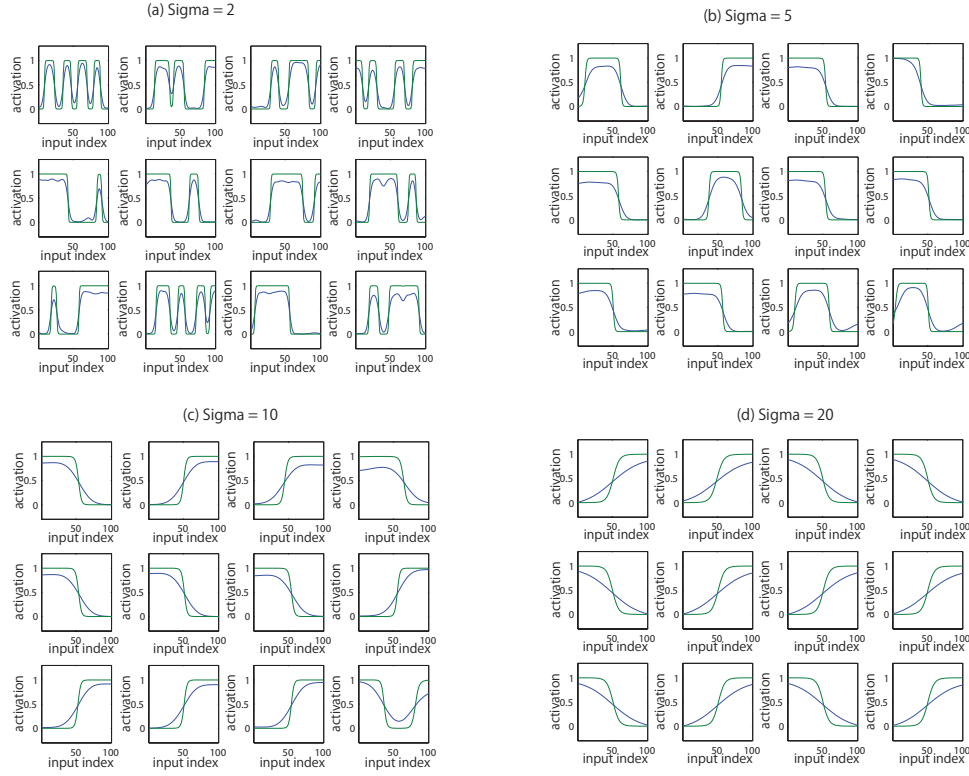
*Figure 22.* Simulation results investigating the development of monotonic tuning responses in the simplified network model with one layer of synapses as described in Appendix A. The current location in the feature space is represented by the position of a Gaussian packet of activity imposed on the 1-dimensional layer of 100 input neurons. The results for three different simulations are shown in separate rows. Top row: network trained with circularly arranged input neurons with wrap-around. Middle row: network trained with linearly arranged input neurons with no wrap-around (divisive competition). Bottom row: network trained with linearly arranged input neurons with no wrap-around (combined lateral inhibition and excitation). The columns show the following: (a) the width,  $\sigma$ , of the Gaussian activity packet imposed on the input layer during training and testing, (b) matrix of synaptic weights from input neurons (ordinate) to output neurons (abscissa), (c) matrix showing activations of output neurons (abscissa) as a Gaussian activity packet is shifted through successive locations on the input layer (ordinate), and (d) matrix showing firing rates of output neurons (abscissa) as a Gaussian activity packet is shifted through the input layer (ordinate). Inspection of these plots shows that the trained networks with linearly arranged input neurons with no wrap-around (middle and bottom rows) have developed output neurons with monotonic tuning responses to the location of the Gaussian activity packet in the input layer regardless of the type of competition. However, output neurons did not show monotonic responses in the network trained with circularly arranged input neurons with wrap-around (top row).



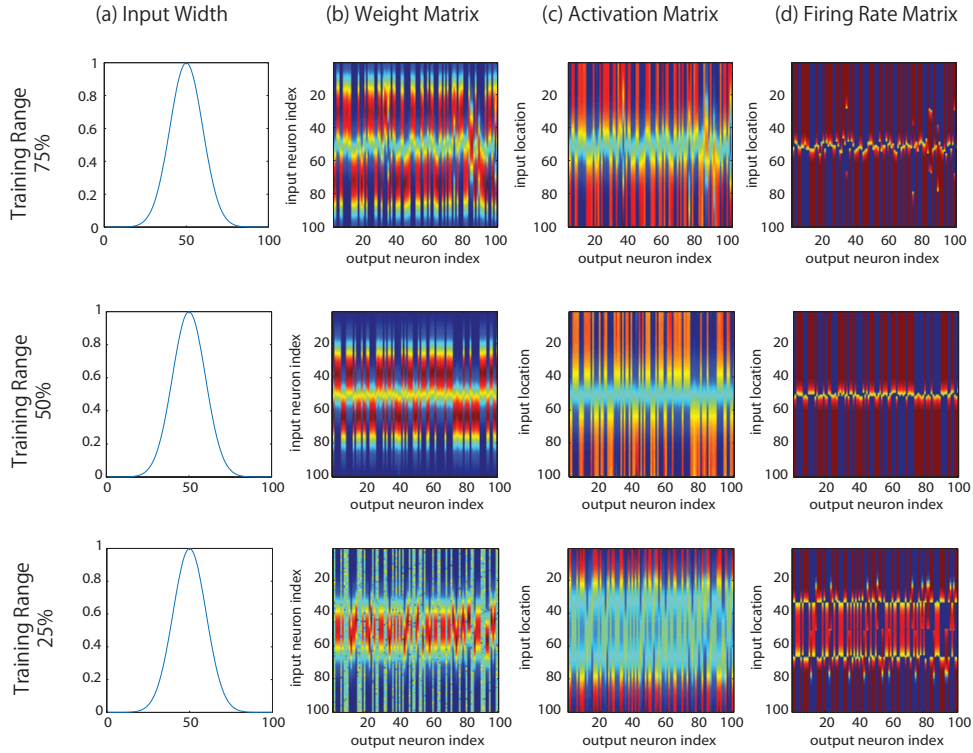
*Figure 23.* Simulation results investigating the development of monotonic tuning responses in the simplified network model with one layer of synapses as described in Appendix A. The results for three different simulations are shown in separate blocks. Block a: network trained with circularly arranged input neurons with wrap-around. Block b: network trained with linearly arranged input neurons with no wrap-around (divisive competition). Block c: network trained with linearly arranged input neurons with no wrap-around (combined lateral inhibition and excitation). For each of the three simulations, we show the behaviour of twelve typical output neurons in separate subplots. Each subplot shows how the activation (blue) and firing rate (green) of the neuron (ordinate) vary as a Gaussian activity packet is shifted through successive locations on the input layer (abscissa). It is evident that regardless of the type of competition implemented, the trained networks with linearly arranged input neurons with no wrap-around (block b and c) have developed output neurons with monotonic tuning responses to the location of the Gaussian activity packet in the input layer. However, this was not the case for output neurons in the network trained with circularly arranged input neurons with wrap-around (block a).



*Figure 24.* Simulation results investigating the development of monotonic tuning responses after training in the simplified network model with one layer of synapses. These simulations implemented divisive inhibition. Four simulations were run with different widths,  $\sigma$ , for the Gaussian packet of activity imposed on the 1-dimensional layer of 100 input neurons. The results for the four different simulations are shown in separate rows with  $\sigma$  set to 2, 5, 10 and 20 neurons. The columns follow the same conventions as in 22 and show the following: (a) the width of the Gaussian input packet, (b) synaptic weight matrix, (c) activation matrix, and (d) firing rate matrix. It can be seen that for a relatively small value of sigma equal to 2, the output neurons do not develop monotonic tuning responses. However, for a large value of  $\sigma = 20$  neurons, the output neurons do display monotonic tuning profiles after training. Thus, the output neurons gradually switch to monotonic tuning curves with increases in the width,  $\sigma$ , of the Gaussian input packet.

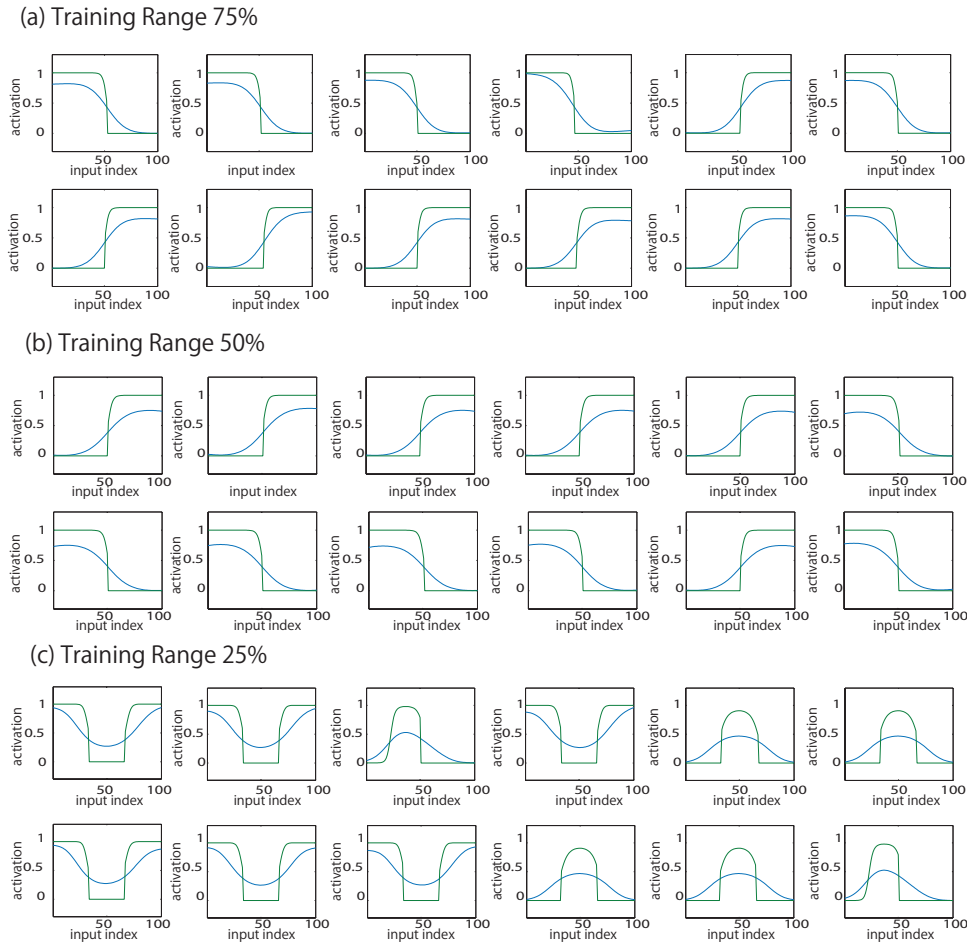


*Figure 25.* Simulation results investigating the development of monotonic tuning responses after training in the simplified network model with one layer of synapses. These simulations implemented divisive inhibition. Four simulations were run with different widths,  $\sigma$ , for the Gaussian packet of activity imposed on the 1-dimensional input layer. The results for the four different simulations are shown in separate blocks: (a)  $\sigma = 2$  neurons, (b)  $\sigma = 5$  neurons, (c)  $\sigma = 10$  neurons, and (d)  $\sigma = 20$  neurons. For each of the four simulations, we show the behaviour of twelve typical output neurons in separate subplots. Each subplot shows how the activation (blue) and firing rate (green) of the neuron (ordinate) vary as a Gaussian activity packet is shifted through successive locations on the input layer (abscissa). It is evident that the output neurons gradually transition to developing monotonic responses as the width,  $\sigma$ , of the Gaussian input packet increases.

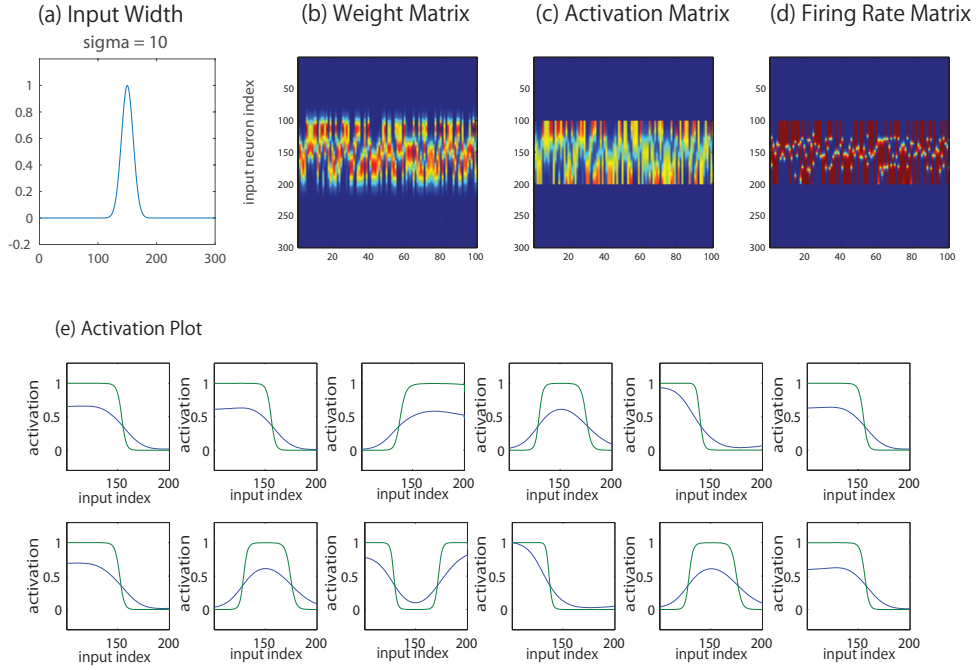


*Figure 26.* Simulation results investigating the development of monotonic tuning responses after training in the simplified network model with one layer of synapses. These simulations implemented divisive inhibition. Three simulations were run in which the Gaussian activity packet was shifted over different sized intervals of the input layer during training. The results for the three simulations are shown in separate rows. Top row: Gaussian activity packet is shifted during training over 75% of the input layer. Middle row: Gaussian activity packet is shifted over 50% of the input layer. Bottom row: Gaussian activity packet is shifted over 25% of the input layer. However, after training, the network is tested with the Gaussian activity packet presented at all locations on the input layer. The columns follow the same conventions as in Fig. 22 and show the following: (a) the width of the Gaussian input packet, (b) synaptic weight matrix, (c) activation matrix, and (d) firing rate matrix. It can be seen that the output neurons develop monotonic tuning when the Gaussian activity packet is shifted over 50% of the input layer during training (top and second row). However, as the Gaussian activity packet is shifted through less of the input layer during training, the output neurons gradually lose their monotonic responses.

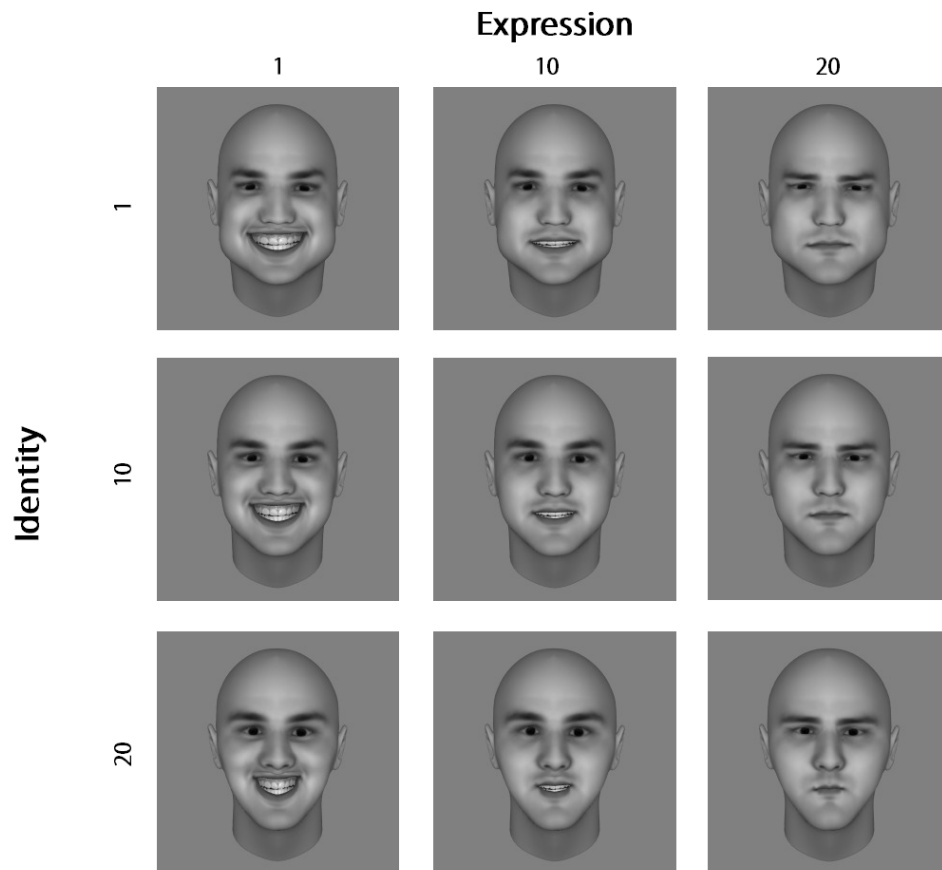




*Figure 27.* Simulation results investigating the development of monotonic tuning responses after training in the simplified network model with one layer of synapses. These simulations implemented divisive inhibition. Three simulations were run in which the Gaussian activity packet was shifted over different sized intervals of the input layer during training. Three simulations were run in which the Gaussian activity packet was shifted over different sized intervals of the input layer during training. The results for the three simulations are shown in separate blocks: (a) Gaussian activity packet is shifted during training over 75% of the input layer, (b) Gaussian activity packet is shifted over 50% of the input layer, (c) Gaussian activity packet is shifted over 25% of the input layer. After training, the network is tested with the Gaussian activity packet presented at all locations on the input layer. For each of the three simulations, we show the behaviour of twelve typical output neurons in separate subplots. Each subplot shows how the activation (blue) and firing rate (green) of the neuron (ordinate) vary as a Gaussian activity packet is shifted through successive locations on the input layer (abscissa). It is evident that the output neurons display monotonic tuning when the Gaussian activity packet has been shifted over 50% of the input layer during training. However, the output neurons gradually lose their monotonic tuning as the Gaussian activity packet is shifted through a smaller interval of the input layer during training.

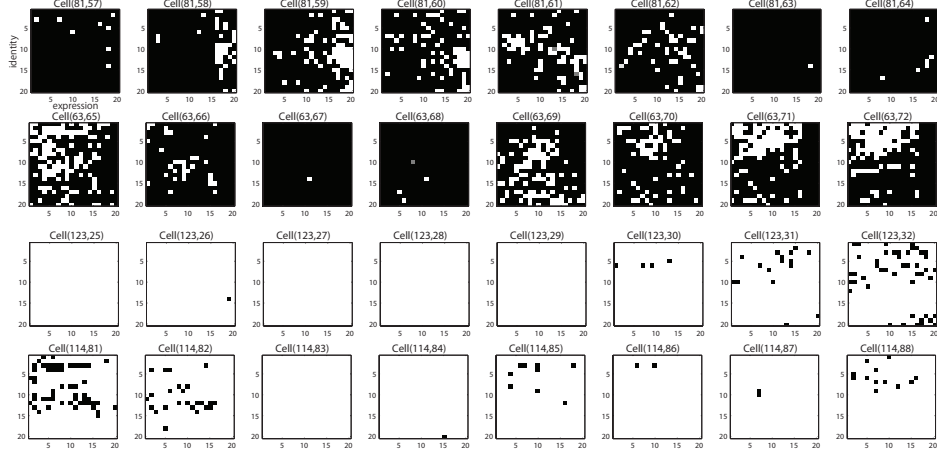


*Figure 28.* Simulation results investigating the development of monotonic tuning responses after training in the simplified network model with one layer of synapses. These simulations implemented divisive inhibition. In this simulation, we extended the input layer to include 100 extra neurons on either side of the 100 original input neurons to ensure that the Gaussian activity packet was not truncated at the original end locations. The columns follow the same conventions as in Fig. 22 and show the following: (a) the width of the Gaussian input packet, (b) synaptic weight matrix, (c) activation matrix, and (d) firing rate matrix. (e) In addition, we show the behaviour of twelve typical output neurons in separate subplots. Each subplot shows how the activation (blue) and firing rate (green) of the neuron (ordinate) vary as a Gaussian activity packet is shifted through successive locations on the input layer (abscissa). It can be seen that some of the neurons show non-monotonic peaked (Gaussian) tuning profiles.

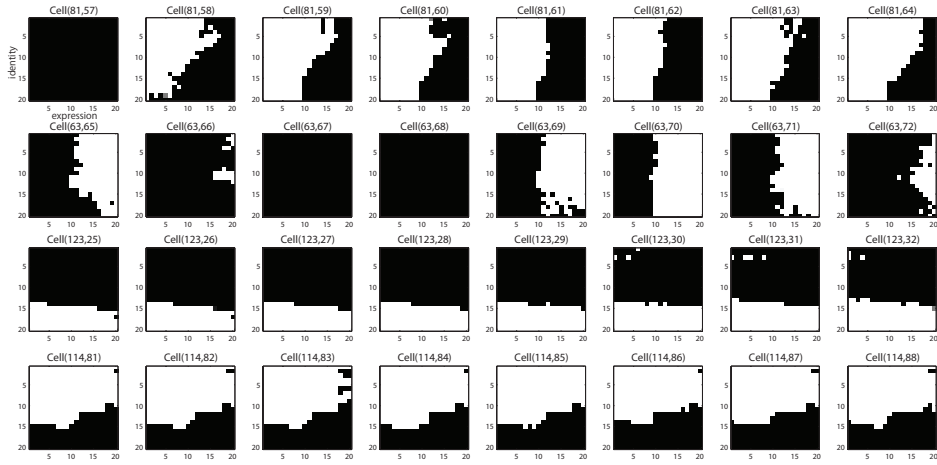


*Figure 29.* The face stimuli used to test VisNet for the existence of neurons that respond selectively to either facial identity or expression. A 1-dimensional space of 20 different facial identities, which varied gradually from one Identity A to another Identity B, was constructed. Each of these identities was then varied over a 1-dimensional space of 20 different expressions from Sad to Happy. This produced a matrix of 400 face stimuli constructed from 20 identities  $\times$  20 expressions.

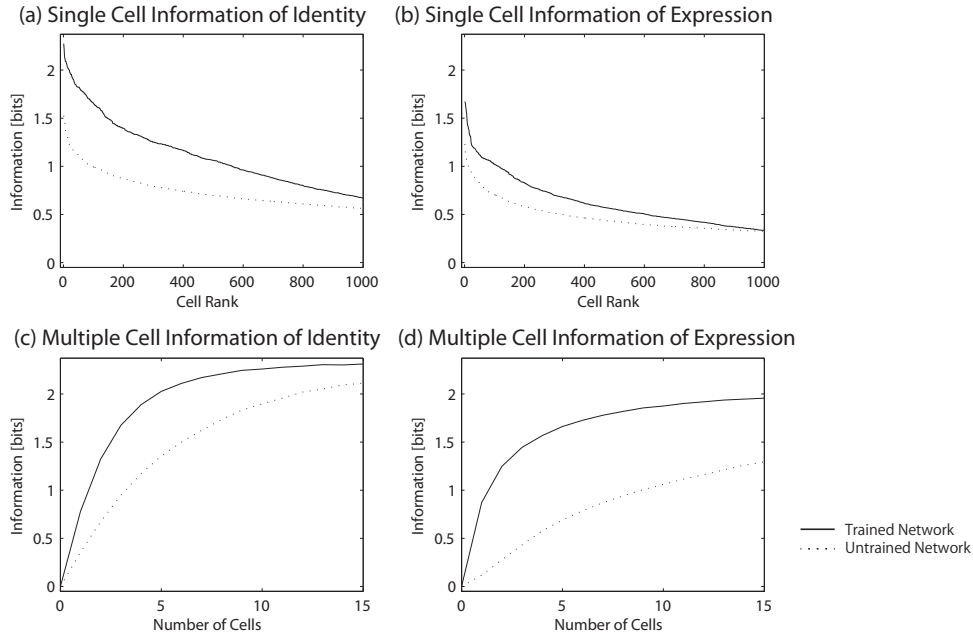
(a) Untrained Network



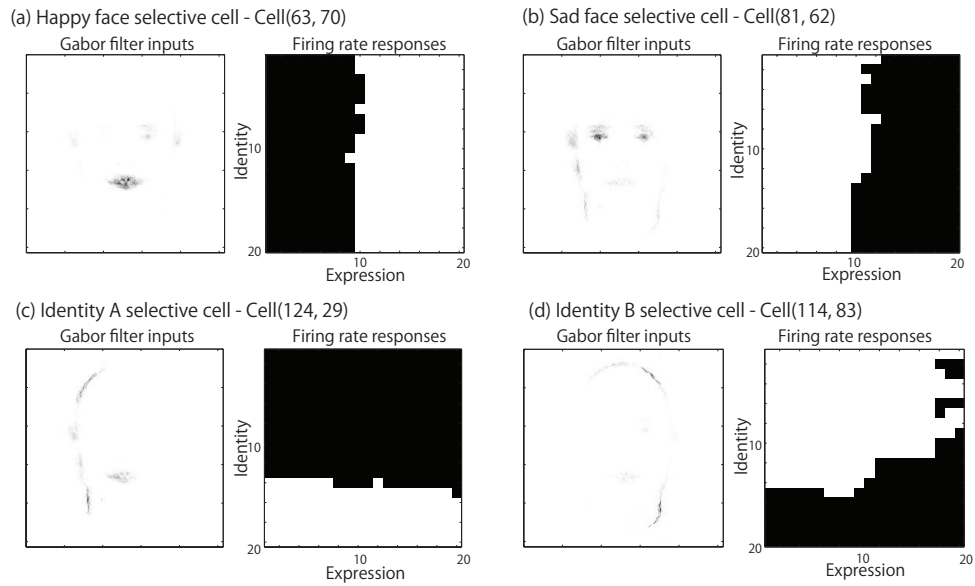
(b) Trained Network



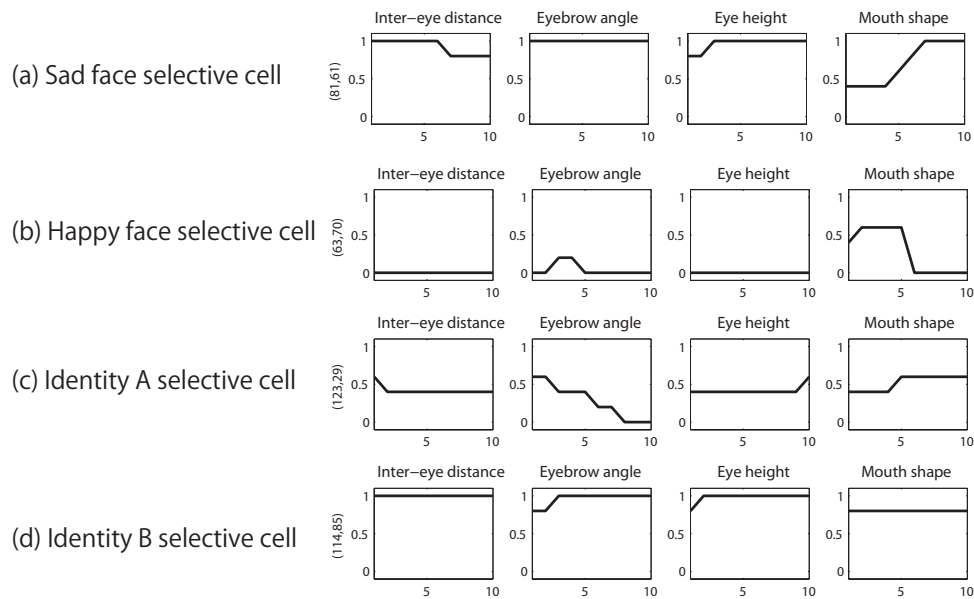
*Figure 30.* Firing rate responses of typical neurons in VisNet when tested on the facial stimuli representing combinations of identity and expression shown in Figure 29. Results are shown before training (a) and after training (b). Each row shows a different block of eight neurons in the 4th layer of the network. For each neuron, we plot its firing rate as a function of facial expression (abscissa) and identity (ordinate), with high firing denoted by black. Before training, the neuronal responses are quite unstructured with respect to facial identity and expression. However, after training, individual neurons respond selectively to localised regions of either the space of identities or space of expressions.



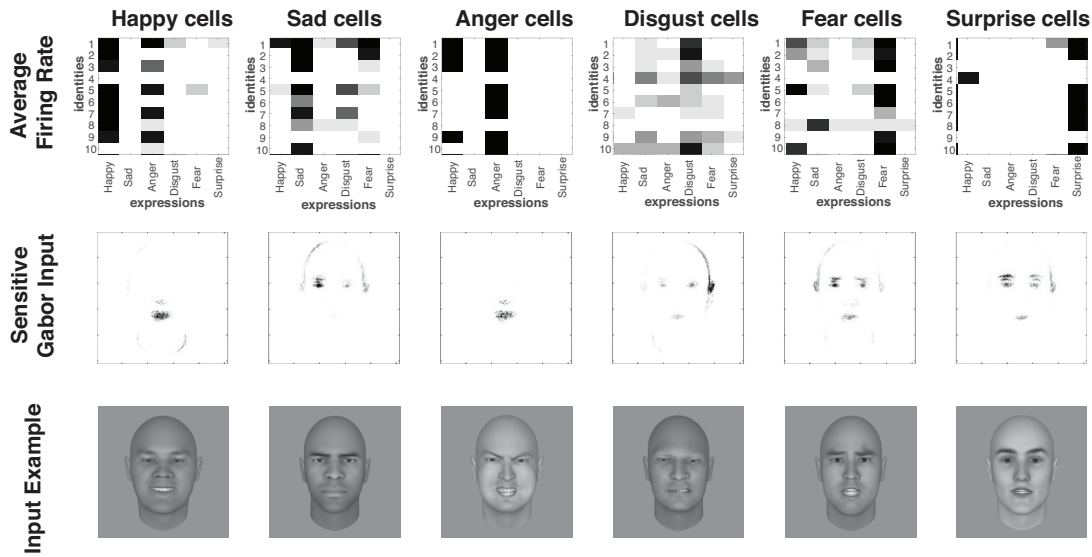
*Figure 31.* Analysis of the amount of single and multiple cell information carried by 4th layer neurons in VisNet about facial identity and expression before and after training. The left column shows the results of analysing the amount of single cell information (a) and multiple cell information (c) about identity conveyed by fourth layer cells before and after training. (a) shows the amount of single cell information carried by output cells plotted in rank order. The dotted line represents the untrained network while the solid line represents the trained network. This analysis involved quantising the identity space into five separate contiguous blocks. The maximal amount of information possible in this case is  $\log_2(5) = 2.32$  bits. The right column shows equivalent results of analysing the amount of single cell information (b) and multiple cell information (d) about expression conveyed by fourth layer cells before and after training. This analysis similarly involved quantising the expression space into five separate contiguous blocks. The maximal amount of information possible is again 2.32 bits. It is evident that training has significantly increased the amount of single and multiple cell information carried by 4th layer neurons about both facial identity and expression.



*Figure 32.* The input Gabor filters that have strong connectivity through the network to example 4th layer neurons in VisNet which are tuned to four global attributes: Happy (top left), Sad (top right), Identity A (bottom left), and Identity B (bottom right). Each of these four subplots shows the Gabor filters with strong connectivity to that neuron (left) and the neuron's firing rate responses to the facial stimuli representing combinations of identity and expression shown in Figure 29.

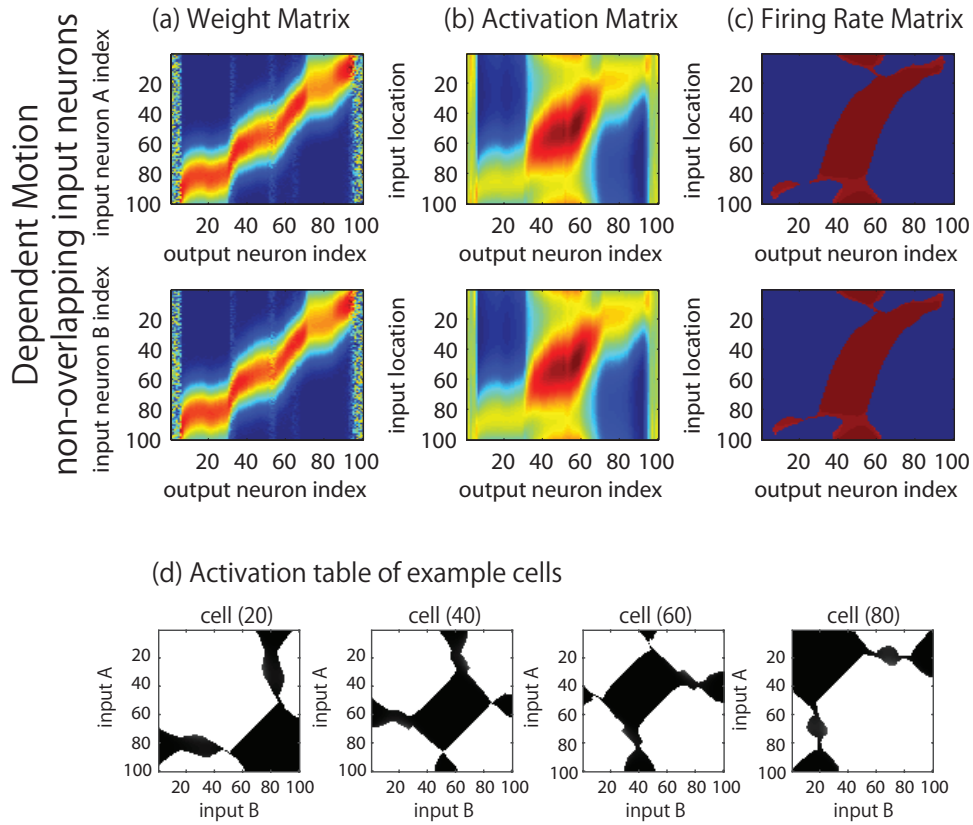


*Figure 33.* How the firing rate responses of four example neurons in the 4th layer of VisNet, which respond selectively to global attributes such as sad, happy, Identity A, and Identity B, depend on variation in the spatial relationships between facial features. Each row shows the responses of a different neuron, while each column corresponds to a different kind of spatial relationship: inter-eye distance, eyebrow angle, eye height and mouth shape. The individual subplots show the firing rate responses of the neuron as the corresponding spatial relationship is varied across ten selected feature values. It is evident that some neurons responding to global attributes such as sad and Identity A have monotonic tuning to particular spatial relationships between local facial features.

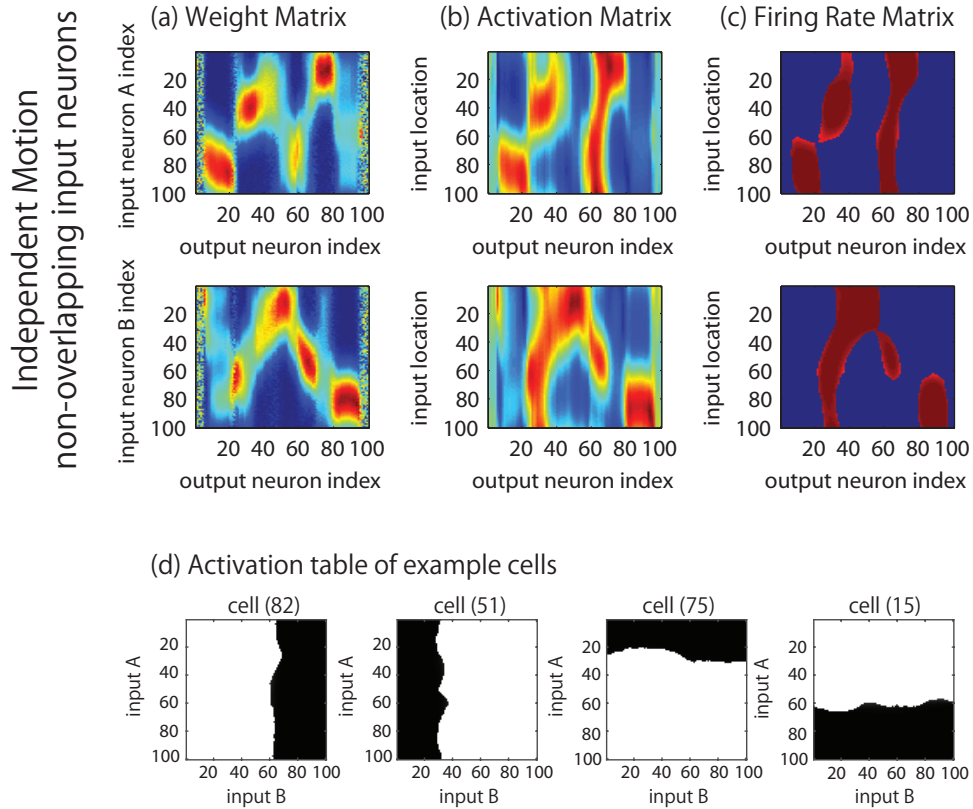


*Figure 34.* Results of the additional VisNet study where the network was trained on a set of 100 randomly generated facial identities for each one of six basic expressions: happy, sad, angry, disgust, fear, and surprise. The network was tested on face stimuli with the same six expressions. For each expression, we created 10 different randomly generated facial identities in order to test whether the network representation of facial expression could generalise across the different facial identities. Each subplot in the top row shows the average firing rate of the ten 4th layer neurons that carry the most information about one of the six expressions, with the neurons encoding each expression shown in a separate column. Each subplot shows the average responses of the ten neurons to ten different randomly generated facial identities with that particular expression. The subplots in the middle row show the average inputs from the gabor filters that are most strongly connected to the ten output cells that represent each expression shown in the top row. The images in the bottom show examples of the randomised face stimuli with the corresponding facial expression used to test the network.

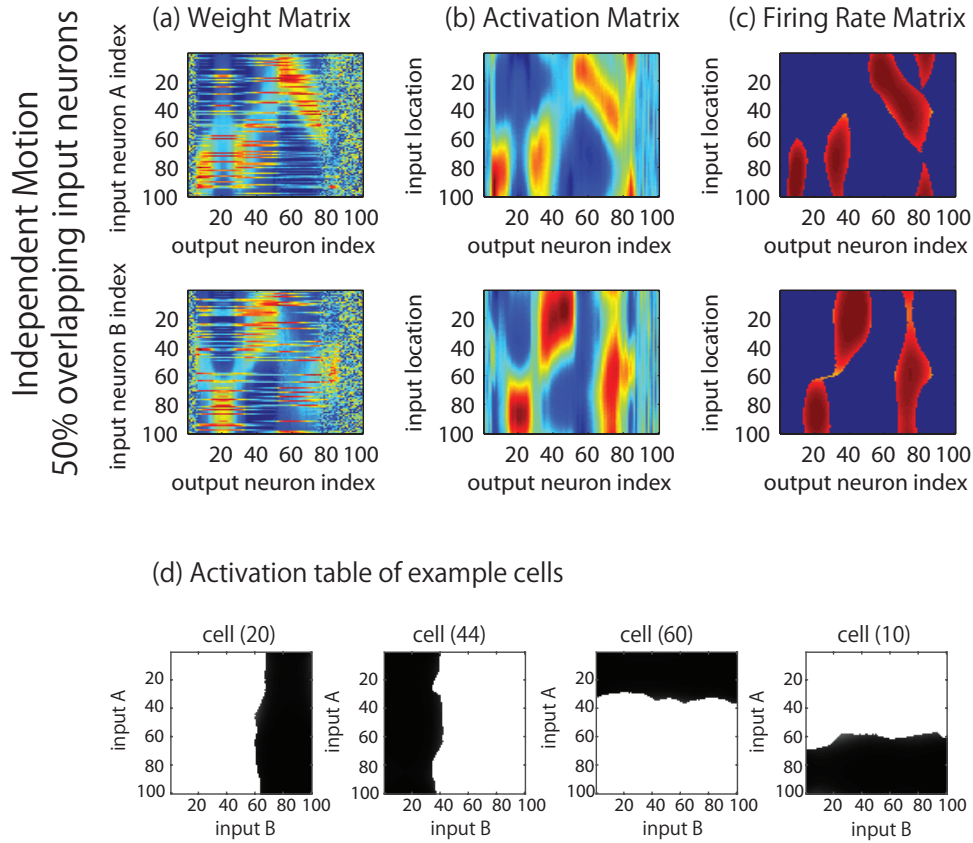




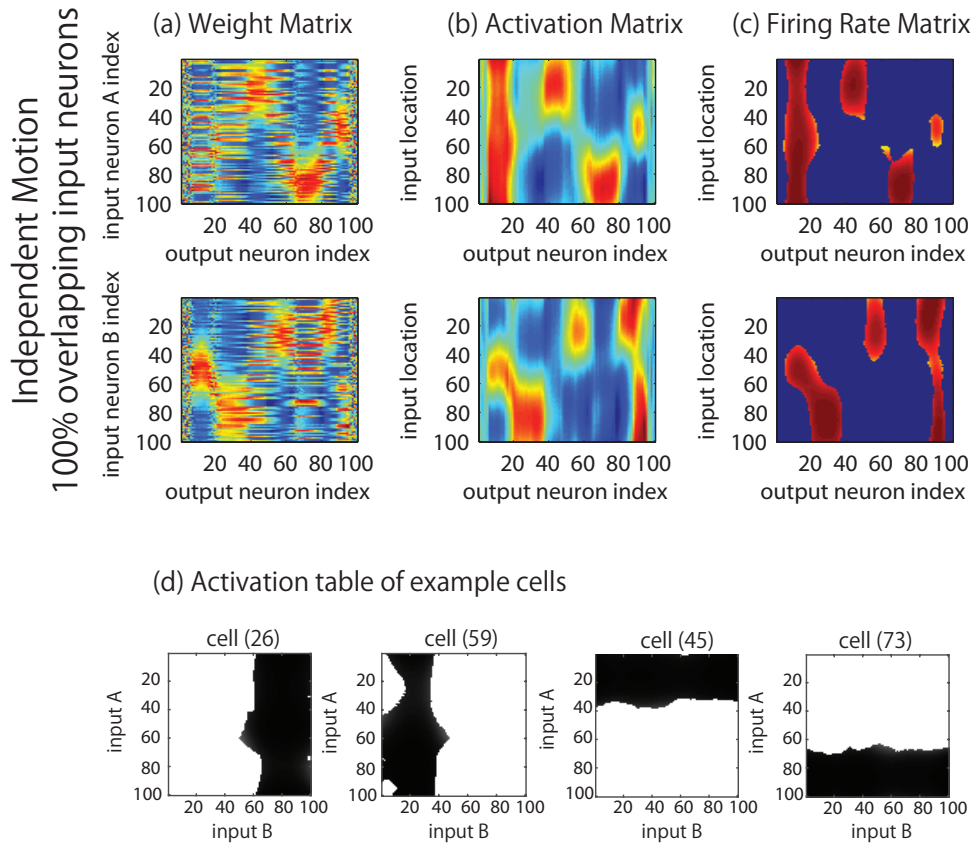
*Figure 35.* Simulation of a one-layer competitive network showing how the *dependent* motion of Gaussian activity packets in two input layers during training affects the learned response properties of output neurons. In this simulation there was no overlap between the two input spaces. The top six subplots show the results of the first method of analysis, in which the position of the Gaussian activity packet in one of the input spaces was systematically shifted through neurons 1 to 100, while the position of the Gaussian packet in the other input space remain fixed at the centre of that space. The first (top) row corresponds to shifting the activity packet through the first input space, while the second row corresponds to shifting the activity packet through the second input space. The three columns show the (a) weight matrix, (b) activation matrix, and (c) firing rate matrix of the population of output neurons. It can be seen that, with dependent motion of the activity packets in the two input layers during training, the output neurons have failed to develop separate representations of the two input spaces. The four subplots in the bottom row (d) show the second method of analysis, in which Gaussian activity packets were presented at all  $100 \times 100$  combinations of positions within the two input spaces, and the firing rate response tables of each output cell were recorded. Each of the four subplots in the bottom row shows the table of firing rate responses for a different output neuron. It is evident that individual output neurons have learned to respond to particular combinations of locations in the two input spaces that occurred together during training.



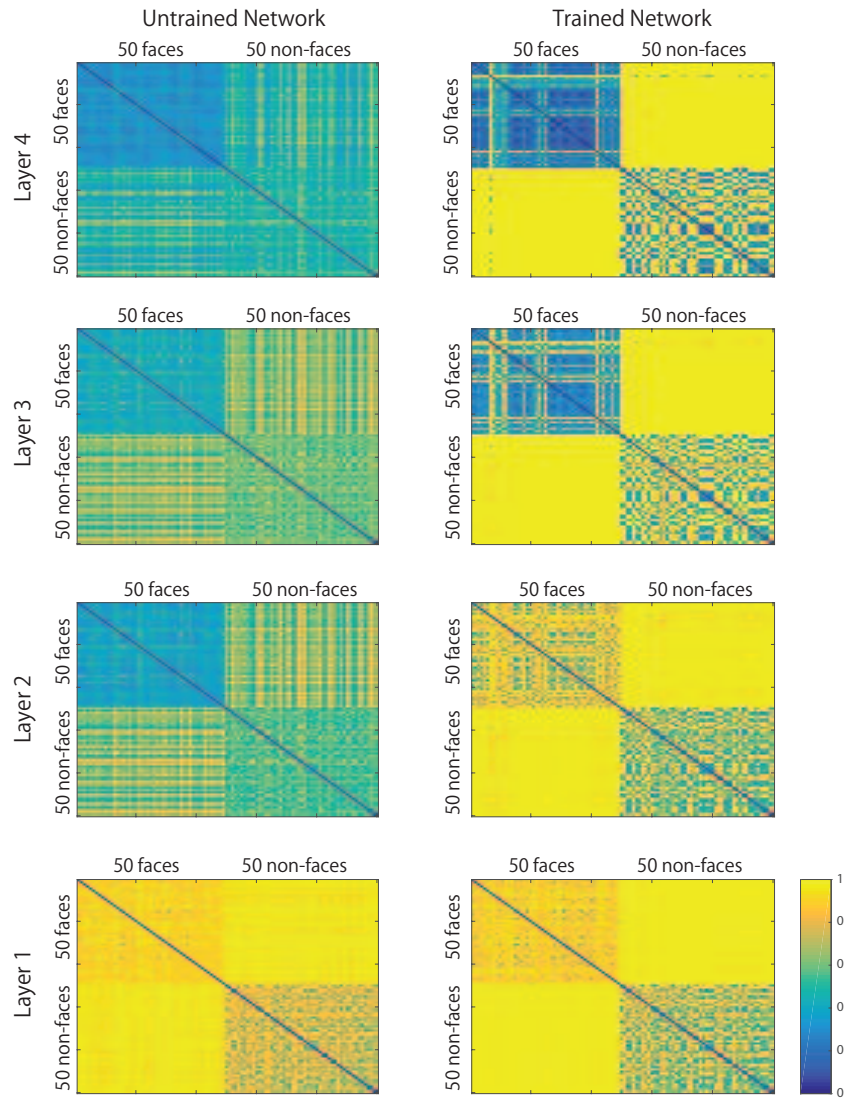
*Figure 36.* Simulation of a one-layer competitive network showing how the *independent* motion of Gaussian activity packets in two input layers during training affects the learned response properties of output neurons. In this simulation there was no overlap between the two input spaces. Conventions as in Figure 35. It can be seen in (a), (b) and (c) that the output neurons have developed separate representations of the two input spaces, with individual neurons responding to just one of the input spaces. In particular, the four output neurons shown in (d) have each learned to respond selectively to a localised end region of one of the input spaces.



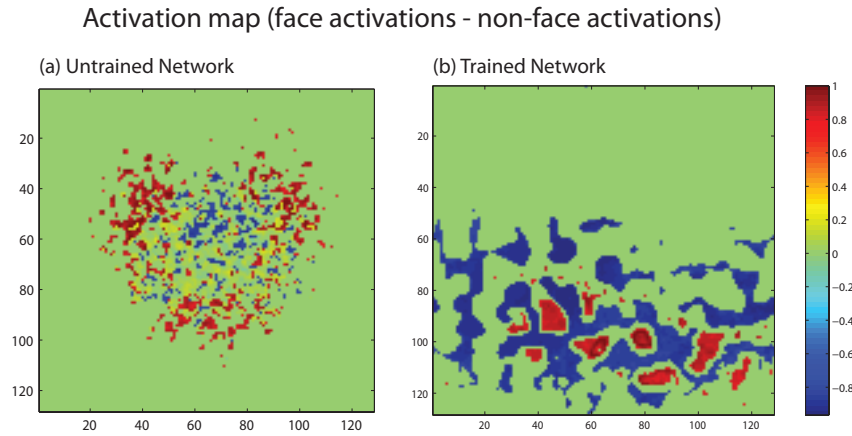
*Figure 37.* Simulation of a one-layer competitive network showing how the *independent* motion of Gaussian activity packets in two input layers during training affects the learned response properties of output neurons. In this simulation there was a 50% overlap between the two input spaces. Conventions as in Figure 35. Even though there is a 50% overlap between the two input spaces, subplots (a), (b) and (c) show that the output neurons still developed separate representations of the two input spaces. The four output neurons shown in (d) have each learned to respond selectively to a localised end region of one of the input spaces.



*Figure 38.* Simulation of a one-layer competitive network showing how the *independent* motion of Gaussian activity packets in two input layers during training affects the learned response properties of output neurons. In this simulation there was a 100% overlap between the two input spaces. That is, exactly the same set of input neurons was used to encode both of the two input spaces. Conventions as in Figure 35. Even though there is a 100% overlap between the two input spaces, subplots (a), (b) and (c) show that the output neurons still developed separate representations of the two input spaces. The four output neurons shown in (d) have each learned to respond selectively to a localised end region of one of the input spaces.



*Figure 39.* Representational dissimilarity matrices (RDM) (Kriegeskorte et al., 2008a) showing the correlations in network activity in response to 50 faces (Figure 11) and 50 non-faces (Figure 12) for each layer of VisNet before training (left column) and after training (right column). For each layer, we recorded the responses of all  $128 \times 128$  neurons in response to each of the 100 test images. We then computed the Pearson correlations between the vectors of neuronal responses across the layer to each pair of test images. A representational dissimilarity matrix was then constructed for each layer where each element corresponding to a particular pair of test images was computed as  $1 - \text{the Pearson correlation}$ . These results show that, after training, the output (4th) layer of the network demonstrates neuronal activity patterns that are highly correlated in response to pairs of stimuli from within one of the stimulus categories, i.e. faces or non-face objects, but are decorrelated in response to stimuli from different categories. It can be seen that this effect gradually increases through successive neuronal layers of the network.



*Figure 40.* Map showing stimulus selectivity of all 4th layer neurons to the faces and non-face objects before training (a) and after training (b). The selectivity measure was computed for all cells that had an average firing rate response greater than or equal to 0.8 for at least one of the stimulus categories. The selectivity measure was calculated by subtracting the average firing rate response of each cell to the non-face objects from the average firing rate response to the faces. The selectivity measure is near +1 (red) for a cell that is selective to faces and near -1 (blue) for a cell that is selective to non-face objects. The selectivity measure was set to zero for those cells with an average firing rate response below 0.8 for both stimulus categories.

## Appendix A

## Model Descriptions

**VisNet Model**

The main simulation studies presented in this paper are conducted with an established biologically plausible neural network model, VisNet, of the primate ventral visual pathway, which was originally developed by Wallis and Rolls (1997). The network architecture is shown in Figure 3. It is based on the following: (i) A series of hierarchical competitive networks with local graded lateral 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 local associative learning rule such as the Hebb rule (6) or trace rule (7), (8), which are explained below.

In past work, the hierarchical series of 4 neuronal layers of VisNet have been related to the following successive stages of processing in the ventral visual pathway: V2, V4, the posterior inferior temporal cortex, and the anterior inferior temporal cortex. However, this correspondence has always been quite loose because the ventral pathway may be further subdivided into a more fine grained network of distinct sub-regions.

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 in the current studies are given in Table 1(a). The gradual increase in the receptive field of cells in successive layers reflects the known physiology of the primate ventral visual pathway (Freeman and Simoncelli, 2011; Pasupathy, 2006; Pettet and Gilbert, 1992).

**Pre-processing of the visual input by Gabor filters.** Before the visual images are presented to the VisNet's input layer 1, they are preprocessed by a set of Gabor filters that accord with the general tuning profiles of simple cells in V1 (Jones and Palmer, 1987; Cumming and Parker, 1999; Lades et al., 1993). The filters provide a unique pattern of filter outputs for each transform of each visual object, which is passed

through to the first layer of VisNet. These filters are known to provide a good fit to the firing properties of V1 simple cells, which respond to local oriented bars and edges within the visual field (Jones and Palmer, 1987; Cumming and Parker, 1999). The input filters used are computed by the following equations:

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

with the following definitions:

$$\begin{aligned} x' &= x \cos \theta + y \sin \theta \\ y' &= -x \sin \theta + y \cos \theta \\ \sigma &= \frac{\lambda(2^b+1)}{\pi(2^b-1)} \sqrt{\frac{\ln 2}{2}} \end{aligned} \quad (2)$$

where  $x$  and  $y$  specify the position of a light impulse in the visual field (Petkov and Kruizinga, 1997). The parameter  $\lambda$  is the wavelength ( $1/\lambda$  is the spatial frequency),  $\sigma$  controls number of such periods inside the Gaussian window based on  $\lambda$  and spatial bandwidth  $b$ ,  $\theta$  defines the orientation of the feature,  $\psi$  defines the phase, and  $\gamma$  sets the aspect ratio that determines the shape of the receptive field. In the experiments in this paper, an array of Gabor filters is generated at each of  $256 \times 256$  retinal locations with the parameters given in Table 1(a). These parameters were selected based on those that previously optimised performance (Tromans et al., 2011; Rolls and Milward, 2000).

The outputs of the Gabor filters are passed to the neurons in layer 1 of VisNet according to the synaptic connectivity given in Table 1(a). That is, each layer 1 neuron receives connections from 201 randomly chosen Gabor filters localised within a topologically corresponding region of the retina.

**Activations of neurons and competition within the network.** Within each of the neural layers 1 to 4 of the network, the activation  $h_i$  of each neuron  $i$  is set equal to a linear sum of the inputs  $r_j$  from afferent neurons  $j$  in the preceding layer



weighted by the synaptic weights  $w_{ij}$ . That is,

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

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

In this paper, we have run simulations with a self-organising map (SOM) (von der Malsburg, 1973; Kohonen, 1982) implemented within each layer. In the SOM architecture, short-range excitation and long-range inhibition are combined to form a Mexican-hat spatial profile and is constructed as a difference of two Gaussians as follows:

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

Here, to implement the SOM, the activations  $h_i$  of neurons within a layer are convolved with a spatial filter,  $I_{ab}$ , where  $\delta_I$  controls the inhibitory contrast and  $\delta_E$  controls the excitatory contrast. The width of the inhibitory radius is controlled by  $\sigma_I$  while the width of the excitatory radius is controlled by  $\sigma_E$ . The parameters  $a$  and  $b$  index the distance away from the centre of the filter. The lateral inhibition and excitation parameters used in the SOM architecture are given in Table 1(a), which were selected based on those that previously optimized performance (Rolls, 2000; Tromans et al., 2011).

Next, the contrast between the activities of neurons with each layer is enhanced by passing the activations of the neurons through a sigmoid transfer function as follows:

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

where  $h'$  is the activation after applying the SOM filter,  $r$  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 within each layer of neurons to control the sparseness of the firing rates. For example, to set

the sparseness to 5%, the threshold is set to the value of the 95th percentile point of the activations within the layer. The parameters for the sigmoid activation function are shown in Table 1(a). They are similar to the standard VisNet sigmoid parameter values that were previously optimised to provide reliable performance (Stringer et al., 2006, 2007; Stringer and Rolls, 2008).

**Modification of synaptic weights during training.** During training with visual objects, the strengths of the feed-forward synaptic connections between successive neuronal layers are modified by biologically plausible local learning rules, where the change in the strength of a synapse depends on the current or recent activities of the pre- and post-synaptic neurons. A variety of such learning rules may be implemented with different learning properties.

One simple well known learning rule is the Hebb rule:

$$\delta w_{ij} = k r_i^\tau r_j^\tau \quad (6)$$

where  $\delta w_{ij}$  is the change of synaptic weight  $w_{ij}$  from pre-synaptic neuron  $j$  to post-synaptic neuron  $i$ ,  $r_i^\tau$  is the firing rate of post-synaptic neuron  $i$  at timestep  $\tau$ ,  $r_j^\tau$  is the firing rate of pre-synaptic neuron  $j$  at timestep  $\tau$ , and  $k$  is the learning rate constant.

Alternatively, a trace learning rule (Foldiak, 1991; Wallis and Rolls, 1997) may be implemented, which incorporates a memory trace of recent neuronal activity:

$$\delta w_{ij} = k \bar{r}_i^{\tau-1} r_j^\tau \quad (7)$$

where  $\bar{r}_i^\tau$  is the trace value of the firing rate of post-synaptic neuron  $i$  at timestep  $\tau$ . The trace term is updated at each timestep according to

$$\bar{r}_i^\tau = (1 - \eta) \bar{r}_i^{\tau-1} + \eta r_i^\tau \quad (8)$$

where  $\eta$  may be set anywhere in the interval  $[0, 1]$ , and for the simulations described

below,  $\eta$  was set to 0.8. The effect of this learning rule is to encourage neurons to learn to respond to visual input patterns that tend to occur close together in time.

To prevent the same few neurons always winning the competition, the synaptic weight vectors are normalised to unit length after each learning update for each training pattern by setting

$$\mathbf{w}_i = \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|} \quad (9)$$

where  $\|\mathbf{w}_i\|$  is the length of the vector  $\mathbf{w}_i$  given by

$$\|\mathbf{w}_i\| = \sqrt{\sum_j w_{ij}^2} \quad (10)$$

Neurophysiological evidence for synaptic weight normalization is provided by Royer and Pare (2003).

### Simplified network model

In order to analyse the learning mechanisms in greater detail, some complementary simulations were also carried out within a much simpler competitive neural network architecture with only one layer of fully connected, associatively modifiable synapses as shown in Figure 9. The network was trained and tested on 1-dimensional Gaussian input patterns, which provided an idealised representation of a 1-dimensional facial feature space such as the distance between the eyes. This abstracted neural network model allowed a more controlled investigation of the hypothesised mechanisms underpinning the development of the cell response characteristics of interest.

**Firing rates of neurons in the input layer.** The population of input neurons represent the current position  $x$  within a 1-dimensional feature space, such as the distance between the eyes. Each input neuron  $j$  is set to respond maximally to a unique position  $x_j$  in the feature space. The firing rate  $r_j$  of each input neuron  $j$  is determined by a Gaussian distribution positioned at  $x_j$  with standard deviation  $\sigma$  as follows:

$$r_j = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-x_j)^2}{2\sigma^2}} \quad (11)$$

The representation across the population of input cells thus takes the form of a Gaussian packet of activity centred on the current position  $x$  within the feature space.

**Activations of neurons and competition within the network.** Within the output layer of the network, the activation  $h_i$  of each output neuron  $i$  is set equal to a linear sum of the inputs  $r_j$  from afferent neurons  $j$  in the preceding input layer weighted by the synaptic weights  $w_{ij}$  as shown in equation (3).

Next, lateral competition is applied between neurons in the output layer. In the simulations reported below, competition is implemented in one of two possible ways as follows.

In the first competition method, divisive inhibition, the activation  $h_i$  of each output neuron  $i$  is divided by the average activation  $\langle h \rangle$  across all output neurons.

The second type of competition method implements a combination of lateral inhibition and excitation between cells in the output layer in order to effect a self-organising map (SOM). Short-range excitation and long-range inhibition are combined to form a ‘Mexican-hat’ spatial filter, which is constructed as a difference of two Gaussians as follows:

$$I_a = -\delta_I \exp\left(-\frac{a^2}{\sigma_I^2}\right) + \delta_E \exp\left(-\frac{a^2}{\sigma_E^2}\right) \quad (12)$$

To implement the SOM, the activations  $h_i$  of neurons within the output layer are convolved with the spatial filter,  $I_a$ , where  $\delta_I$  controls the inhibitory contrast and  $\delta_E$  controls the excitatory contrast. The width of the inhibitory radius is controlled by  $\sigma_I$  while the width of the excitatory radius is controlled by  $\sigma_E$ . The parameter  $a$  indexes the distance away from the centre of the filter. The lateral inhibition and excitation parameters are given in Table 1(b).

Next, the contrast between the activities of neurons with the output layer is enhanced by passing the activation  $h_i$  of each output neuron,  $i$ , through a sigmoid transfer function as shown in equation (5). The sigmoid slope  $\beta$  is set to a fixed value throughout each simulation given in Table 1(b). However, the sigmoid threshold  $\alpha$  is continually adjusted to control the sparseness of the firing-rates within the output layer.

**Modification of synaptic weights during training.** At the beginning of each simulation, the synaptic weights from the input neurons to the output neurons are initialized with random values.

The simulation begins with a training phase. At each timestep  $\tau$  during training, the firing rates of the input neurons are first updated to represent a new position  $x$  in the feature space, and then the firing rates of the output neurons are computed as described above. Then the synaptic weights are updated according to an associative (hebbian) learning rule as described in equation (6)

To prevent the same few output neurons always winning the competition, the synaptic weight vector  $\mathbf{w}_i$  of each output neuron  $i$  is renormalised after each learning update by equation (9) and (10).

## Appendix B

## Information Analysis

To quantify the performance in transformation invariance learning with VisNet, the techniques of Shannon’s information theory have previously been used (Rolls and Treves, 1998), which is based on the KL divergence of the conditional response distribution from the unconditional distribution. Information theory can be used to quantify how selective neurons are for particular stimuli, each of which may translate across different locations on the retina. If the responses  $r$  of a neuron carry a high level of information about the presence of a particular stimulus  $s$ , then this implies that the neuron will respond selectively to the presence of that stimulus regardless of where the stimulus is presented on the retina. In this way, information theory can provide a direct measure of both the selectivity of a neuron for a particular stimulus, as well as how translation-invariant the neuronal responses are as the stimulus is shifted across the retina.

Two information measures were used to assess the ability of the network to develop neurons that are selective to the presence of stimuli but also invariant to their occurrence in different retinal locations (see Rolls et al. (1997); Rolls and Milward (2000)). These two measure use the responses from either individual neurons (single-cell information analysis) or small ensembles of neurons (multiple-cell information analysis), each of which will be discussed in turn.

The following exposition provides a theoretical account of the two information measures used in this thesis. However, in order to keep the notation consistent with past publications (Rolls et al., 1997; Rolls and Milward, 2000), we have here denoted the neuronal firing rates by  $r$ .

**Single-cell information**

A single cell information measure was applied to individual cells to measure how much information is available from the responses of a single cell about which color input is present. The amount of stimulus specific information that a certain cell transmits is

calculated from the following formula with details given by Rolls and Milward (2000):

$$I(s, \vec{R}) = \sum_{r \in \vec{R}} P(r|s) \log_2 \frac{P(r|s)}{P(r)} \quad (13)$$

Here  $s$  is a particular stimulus and  $\vec{R}$  is the set of responses of a cell to the set of stimuli. The maximum information that an ideally developed cell could carry is given by the formula:

$$\text{Maximum cell information} = \log_2(n) \text{ bits} \quad (14)$$

where  $n$  is a number of different stimuli.

### Multiple-cell information

While useful in assessing the tuning properties of a particular neuron, the single-cell information measure cannot give a complete assessment of VisNet's performance with respect to recognition of the set of visual stimuli. If all cells learned to respond to the same stimulus (according to the single-cell measure) then there would be relatively little information available about the whole set stimuli  $\vec{S}$ . To address this issue, we also calculated a multiple-cell information measure, which assesses the amount of information that is available about the whole set of the categories of the visual stimuli from a *population* of neurons. This measure quantifies the network's ability to tell which stimulus is currently exposed to the network based on the set of responses,  $\vec{R}$ , of a sub-population of cells. Here we adapt the procedures for calculating the multiple-cell information measure as described by Rolls and Milward (2000); Eguchi et al. (2014).

In brief, we would like to calculate the mutual information between the stimuli and the responses – the average amount of information obtained (across all stimuli) from the responses of the ensemble, about which stimulus was present after a single presentation of a stimulus. However, due to the difficulty in adequately sampling this high dimensional neural response space, it is unrealistic to construct accurate probability distributions for directly calculating the mutual information. Instead, a decoding procedure is used to estimate which stimulus  $s'$  gave rise to the particular

firing rate response vector on each trial. A probability table is then constructed between the real stimuli,  $s$  and the decoded stimuli,  $s'$ . From this probability table, the multiple-cell information is then calculated as follows.

$$I_{\vec{C}}(S, S') = \sum_{s, s'} P(s, s') \log_2 \frac{P(s, s')}{P(s)P(s')} \quad (15)$$

$$P(s') = \sum_{s \in S} P(s' | R_{\vec{C}}(s)) \times P(R_{\vec{C}}(s)) \quad (16)$$

$$P(s, s') = P(s' | R_{\vec{C}}(s)) \times P(R_{\vec{C}}(s)) \quad (17)$$

Here,  $S$  represents the set of the stimuli presented to the networks, and  $\vec{C}$  defines the set of cells used in the analysis, which had as single cells the most information about which stimulus was present. From the set of cells  $\vec{C}$ , the firing responses  $R_{\vec{C}}$  ( $R = r(c) | c \in \vec{C}$ ) to each stimulus in  $S$  are used as the basis for the Bayesian decoding procedure as follows:

$$P(s' | R_{\vec{C}}) = \frac{P(s') \prod_{c \in \vec{C}} P(R_c(s') | s')}{\sum_{s'' \in S} P(s'') \prod_{c \in \vec{C}} P(R_c(s'') | s'')} \quad (18)$$

$$P(R_c(s) | s') = \frac{\sum_{t=1}^{nTrans} pdf(R_c(s, t), \bar{R}_c(s'), SD_c(s'))}{nTrans} \quad (19)$$

where  $nTrans$  defines the number of possible transforms, and  $pdf$  computes the probability density function at firing response of a subset of cells when exposed to a stimulus  $s$  at  $t^{th}$  transforms using the normal distribution with their mean and standard deviation.

For a given set of cells, the probabilities generated by the decoding procedure are factored into a confusion matrix, that matches up the actual input stimuli in  $\vec{S}$  with the predicted stimuli in  $\vec{S}'$ . Here,  $P(s'_i)$  represents the probability that the predicted stimulus  $s'_i$  is actually the stimulus  $s_i$  that is currently presented to the network. A higher value of  $P(s, s')$  relative to  $P(s)P(s')$  indicates a stronger relationship between  $s$  and  $s'$ ; this information provides the basis for calculating the multiple-cell information analysis.



## Appendix C

## Data Sharing

<sup>1647</sup> The VisNet simulator can be downloaded from <https://github.com/bedeho/VisBack>.

## References

1648

- 1649 Baker, C. I., Behrmann, M., and Olson, C. R. (2002). Impact of learning on  
 1650 representation of parts and wholes in monkey inferotemporal cortex. *Nature*  
 1651 *Neuroscience*, 5(11):1210–1216.
- 1652 Beeck, H. P. O. d., Baker, C. I., DiCarlo, J. J., and Kanwisher, N. G. (2006).  
 1653 Discrimination Training Alters Object Representations in Human Extrastriate  
 1654 Cortex. *The Journal of Neuroscience*, 26(50):13025–13036.
- 1655 Benson, P. J. and Perrett, D. I. (1991). Synthesising continuous-tone caricatures. *Image*  
 1656 *and Vision Computing*, 9(2):123–129.
- 1657 Biederman, I. (1987). Recognition-by-components: a theory of human image  
 1658 understanding. *Psychological Review*, 94(2):115–147.
- 1659 Biederman, I. and Kalocsai, P. (1997). Neurocomputational bases of object and face  
 1660 recognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*,  
 1661 352(1358):1203–1219.
- 1662 Bruce, V. and Young, A. (1986). Understanding face recognition. *British Journal of*  
 1663 *Psychology*, 77(3):305–327.
- 1664 Bruce, V. and Young, A. (2011). *Face Perception*. Psychology Press, London ; New  
 1665 York.
- 1666 Calder, A. J., Burton, A. M., Miller, P., Young, A. W., and Akamatsu, S. (2001). A  
 1667 principal component analysis of facial expressions. *Vision Research*, 41(9):1179–1208.
- 1668 Chance, J. E., Turner, A. L., and Goldstein, A. G. (1982). Development of differential  
 1669 recognition for own- and other-race faces. *The Journal of Psychology*, 112(1st  
 1670 Half):29–37.
- 1671 Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of  
 1672 cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204.

- 1673 Cumming, B. G. and Parker, A. J. (1999). Binocular neurons in v1 of awake monkeys  
1674 are selective for absolute, not relative, disparity. *The Journal of Neuroscience*,  
1675 19(13):5602–5618. PMID: 10377367.
- 1676 Eguchi, A., Mender, B. M. W., Evans, B., Humphreys, G., and Stringer, S. (2015).  
1677 Computational modeling of the neural representation of object shape in the primate  
1678 ventral visual system. *Frontiers in Computational Neuroscience*, 9(100).
- 1679 Eguchi, A., Neymotin, S. A., and Stringer, S. M. (2014). Color opponent receptive fields  
1680 self-organize in a biophysical model of visual cortex via spike-timing dependent  
1681 plasticity. *Frontiers in Neural Circuits*, 8(16).
- 1682 Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., and  
1683 Rasmussen, D. (2012). A Large-Scale Model of the Functioning Brain. *Science*,  
1684 338(6111):1202–1205.
- 1685 Engell, A. D. and Haxby, J. V. (2007). Facial expression and gaze-direction in human  
1686 superior temporal sulcus. *Neuropsychologia*, 45(14):3234–3241.
- 1687 Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural*  
1688 *Computation*, 3(2):194–200.
- 1689 Freeman, J. and Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature*  
1690 *Neuroscience*, 14(9):1195–1201.
- 1691 Freeman, J. B., Rule, N. O., Adams, R. B., and Ambady, N. (2010). The neural basis of  
1692 categorical face perception: graded representations of face gender in fusiform and  
1693 orbitofrontal cortices. *Cerebral Cortex (New York, N.Y.: 1991)*, 20(6):1314–1322.
- 1694 Freiwald, W. A., Tsao, D. Y., and Livingstone, M. S. (2009). A face feature space in the  
1695 macaque temporal lobe. *Nature Neuroscience*, 12(9):1187–1196.
- 1696 Friesen, W. V. and Ekman, P. (1976). *Pictures of Facial Affect*. Consulting  
1697 psychologists Press.

- 1698 Gillebert, C. R., Op de Beeck, H. P., Panis, S., and Wagemans, J. (2008). Subordinate  
1699 Categorization Enhances the Neural Selectivity in Human Object-selective Cortex for  
1700 Fine Shape Differences. *Journal of Cognitive Neuroscience*, 21(6):1054–1064.
- 1701 Gosselin, F. and Schyns, P. G. (2001). Bubbles: a technique to reveal the use of  
1702 information in recognition tasks. *Vision Research*, 41(17):2261–2271.
- 1703 Gross, C. G., Rocha-Miranda, C. E., and Bender, D. B. (1972). Visual properties of  
1704 neurons in inferotemporal cortex of the macaque. *Journal of Neurophysiology*,  
1705 35(1):96–111.
- 1706 Hasselmo, M. E., Rolls, E. T., and Baylis, G. C. (1989). The role of expression and  
1707 identity in the face-selective responses of neurons in the temporal visual cortex of the  
1708 monkey. *Behavioural brain research*, 32(3):203–218. PMID: 2713076.
- 1709 Haxby, Hoffman, and Gobbini (2000). The distributed human neural system for face  
1710 perception. *Trends in cognitive sciences*, 4(6):223–233. PMID: 10827445.
- 1711 Henriksson, L., Mur, M., and Kriegeskorte, N. (2015). Faciotopy - a face-feature map  
1712 with face-like topology in the human occipital face area. *Cortex*, 72:156–167.
- 1713 Homola, G. A., Jbabdi, S., Beckmann, C. F., and Bartsch, A. J. (2012). A brain  
1714 network processing the age of faces. *PLoS ONE*, 7(11):e49451.
- 1715 Jones, J. P. and Palmer, L. A. (1987). The two-dimensional spatial structure of simple  
1716 receptive fields in cat striate cortex. *Journal of neurophysiology*, 58(6):1187–1211.  
1717 PMID: 3437330.
- 1718 Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: A  
1719 module in human extrastriate cortex specialized for face perception. *The Journal of*  
1720 *Neuroscience*, 17(11):4302–4311. PMID: 9151747.
- 1721 Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep supervised, but not  
1722 unsupervised, models may explain IT cortical representation. *PLOS Comput Biol*,  
1723 10(11):e1003915.

- 1724 Kiani, R., Esteky, H., Mirpour, K., and Tanaka, K. (2007). Object category structure in  
1725 response patterns of neuronal population in monkey inferior temporal cortex. *Journal*  
1726 *of Neurophysiology*, 97(6):4296–4309.
- 1727 Kohonen, T. (1982). Self-organized formation of topologically correct feature maps.  
1728 *Biological Cybernetics*, 43(1):59–69.
- 1729 Kriegeskorte, N., Mur, M., Bandettini, P. A., Kriegeskorte, N., Mur, M., and  
1730 Bandettini, P. (2008a). Representational similarity analysis - connecting the branches  
1731 of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.
- 1732 Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K.,  
1733 and Bandettini, P. A. (2008b). Matching categorical object representations in inferior  
1734 temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141.
- 1735 Lades, M., Vorbruggen, J., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R.,  
1736 and Konen, W. (1993). Distortion invariant object recognition in the dynamic link  
1737 architecture. *IEEE Transactions on Computers*, 42(3):300–311.
- 1738 Lawrence, S., Giles, C., Tsoi, A. C., and Back, A. (1997). Face recognition: a  
1739 convolutional neural-network approach. *IEEE Transactions on Neural Networks*,  
1740 8(1):98–113.
- 1741 Leder, H. and Bruce, V. (1998). Local and relational aspects of face distinctiveness.  
1742 *The Quarterly Journal of Experimental Psychology. A, Human Experimental*  
1743 *Psychology*, 51(3):449–473.
- 1744 Lisetti, C. L. and Rumelhart, D. E. (1998). Facial Expression Recognition Using a  
1745 Neural Network. In *FLAIRS Conference*, pages 328–332.
- 1746 Mangini, M. C. and Biederman, I. (2004). Making the ineffable explicit: estimating the  
1747 information employed for face classifications. *Cognitive Science*, 28(2):209–226.
- 1748 Masquelier, T. and Thorpe, S. J. (2007). Unsupervised Learning of Visual Features  
1749 through Spike Timing Dependent Plasticity. *PLoS Comput Biol*, 3(2):e31.

- 1750 Maurer, D., Grand, R. L., and Mondloch, C. J. (2002). The many faces of configural  
1751 processing. *Trends in Cognitive Sciences*, 6(6):255–260.
- 1752 Morin, E. L., Hadj-Bouziane, F., Stokes, M., Ungerleider, L. G., and Bell, A. H. (2014).  
1753 Hierarchical encoding of social cues in primate inferior temporal cortex. *Cerebral*  
1754 *cortex* (New York, N.Y.: 1991). PMID: 24836688.
- 1755 Pasupathy, A. (2006). Neural basis of shape representation in the primate brain.  
1756 *Progress in brain research*, 154:293–313. PMID: 17010719.
- 1757 Perrett, D. I., Hietanen, J. K., Oram, M. W., and Benson, P. J. (1992). Organization  
1758 and functions of cells responsive to faces in the temporal cortex. *Philosophical*  
1759 *transactions of the Royal Society of London. Series B, Biological sciences*,  
1760 335(1273):23–30. PMID: 1348133.
- 1761 Petkov, N. and Kruizinga, P. (1997). Computational models of visual neurons  
1762 specialised in the detection of periodic and aperiodic oriented visual stimuli: bar and  
1763 grating cells. *Biological cybernetics*, 76(2):83–96. PMID: 9116079.
- 1764 Pettet, M. W. and Gilbert, C. D. (1992). Dynamic changes in receptive-field size in cat  
1765 primary visual cortex. *Proceedings of the National Academy of Sciences*,  
1766 89(17):8366–8370. PMID: 1518870.
- 1767 Pitcher, D., Walsh, V., and Duchaine, B. (2011). The role of the occipital face area in  
1768 the cortical face perception network. *Experimental Brain Research*, 209(4):481–493.
- 1769 Rhodes, G. (1997). *Superportraits: Caricatures and Recognition*. Psychology Press,  
1770 Hove, East Sussex, UK, 1 edition edition.
- 1771 Rolls, E. T. (2000). Functions of the primate temporal lobe cortical visual areas in  
1772 invariant visual object and face recognition. *Neuron*, 27(2):205–218. PMID: 10985342.
- 1773 Rolls, E. T. and Milward, T. (2000). A model of invariant object recognition in the  
1774 visual system: learning rules, activation functions, lateral inhibition, and

- 1775 information-based performance measures. *Neural computation*, 12(11):2547–2572.  
1776 PMID: 11110127.
- 1777 Rolls, E. T. and Treves, A. (1998). *Neural Networks and Brain Function*. Oxford  
1778 University Press, USA, 1 edition.
- 1779 Rolls, E. T., Treves, A., Tovee, M. J., and Panzeri, S. (1997). Information in the  
1780 neuronal representation of individual stimuli in the primate temporal visual cortex.  
1781 *Journal of computational neuroscience*, 4(4):309–333.
- 1782 Royer, S. and Pare, D. (2003). Conservation of total synaptic weight through balanced  
1783 synaptic depression and potentiation. *Nature*, 422(6931):518–522. PMID: 12673250.
- 1784 Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations  
1785 by back-propagating errors. *Nature*, 323(6088):533–536.
- 1786 Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., and Poggio, T. (2005). *A  
1787 theory of object recognition: computations and circuits in the feedforward path of the  
1788 ventral stream in primate visual cortex*. MIT CSAIL, MA, USA.
- 1789 Sormaz, M., Watson, D. M., Smith, W. A. P., Young, A. W., and Andrews, T. J.  
1790 (2016). Modelling the perceptual similarity of facial expressions from image statistics  
1791 and neural responses. *NeuroImage*, 129:64–71.
- 1792 Stork, D. (1989). Is backpropagation biologically plausible? In , *International Joint  
1793 Conference on Neural Networks, 1989. IJCNN*, pages 241–246 vol.2.
- 1794 Stringer, S. M., Perry, G., Rolls, E. T., and Proske, J. H. (2006). Learning invariant  
1795 object recognition in the visual system with continuous transformations. *Biological  
1796 cybernetics*, 94(2):128–142. PMID: 16369795.
- 1797 Stringer, S. M. and Rolls, E. T. (2008). Learning transform invariant object recognition  
1798 in the visual system with multiple stimuli present during training. *Neural networks:  
1799 the official journal of the International Neural Network Society*, 21(7):888–903.  
1800 PMID: 18440774.

- 1801 Stringer, S. M., Rolls, E. T., and Tromans, J. M. (2007). Invariant object recognition  
1802 with trace learning and multiple stimuli present during training. *Network (Bristol,*  
1803 *England)*, 18(2):161–187. PMID: 17966074.
- 1804 Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). DeepFace: Closing the Gap  
1805 to Human-Level Performance in Face Verification. In *2014 IEEE Conference on*  
1806 *Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708.
- 1807 Tanaka, J. W. and Farah, M. J. (1993). Parts and wholes in face recognition. *The*  
1808 *Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*,  
1809 46(2):225–245.
- 1810 Tromans, J. M. (2012). *Computational neuroscience of natural scene processing in the*  
1811 *ventral visual pathway*. Ph.D., University of Oxford.
- 1812 Tromans, J. M., Harris, M., and Stringer, S. M. (2011). A computational model of the  
1813 development of separate representations of facial identity and expression in the  
1814 primate visual system. *PLoS ONE*, 6(10):e25616.
- 1815 Tromans, J. M., Page, H. J., and Stringer, S. M. (2012). Learning separate visual  
1816 representations of independently rotating objects. *Network: Computation in Neural*  
1817 *Systems*, 23(1-2):1–23.
- 1818 Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., and Livingstone, M. S. (2006). A  
1819 cortical region consisting entirely of face-selective cells. *Science (New York, N.Y.)*,  
1820 311(5761):670–674. PMID: 16456083 PMCID: PMC2678572.
- 1821 von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the  
1822 striate cortex. *Kybernetik*, 14(2):85–100.
- 1823 von der Malsburg, C. (1981). The Correlation Theory of Brain Function. Departmental  
1824 Technical Report, MPI.
- 1825 von der Malsburg, C. and Schneider, W. (1986). A neural cocktail-party processor.  
1826 *Biological Cybernetics*, 54(1):29–40.



- 1827 Wallis, G. (2013). Toward a unified model of face and object recognition in the human  
1828 visual system. *Frontiers in Psychology*, 4:497.
- 1829 Wallis, G. and Rolls, E. T. (1997). Invariant face and object recognition in the visual  
1830 system. *Progress in Neurobiology*, 51(2):167–194.
- 1831 Wegrzyn, M., Riehle, M., Labudda, K., Woermann, F., Baumgartner, F., Pollmann, S.,  
1832 Bien, C. G., and Kissler, J. (2015). Investigating the brain basis of facial expression  
1833 perception using multi-voxel pattern analysis. *Cortex*, 69:131–140.
- 1834 Xu, X., Biederman, I., and Shah, M. P. (2014). A neurocomputational account of the  
1835 face configural effect. *Journal of Vision*, 14(8):9.
- 1836 Yankouskaya, A., Humphreys, G. W., and Rotshtein, P. (2014). Differential interactions  
1837 between identity and emotional expression in own and other-race faces: effects of  
1838 familiarity revealed through redundancy gains. *Journal of Experimental Psychology.*  
1839 *Learning, Memory, and Cognition*, 40(4):1025–1038.
- 1840 Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*,  
1841 81(1):141–145.
- 1842 Yue, X., Tjan, B. S., and Biederman, I. (2006). What makes faces special? *Vision*  
1843 *research*, 46(22):3802–3811.
- 1844 Zhang, J., Li, X., Song, Y., and Liu, J. (2012). The fusiform face area is engaged in  
1845 holistic, not parts-based, representation of faces. *PLoS ONE*, 7(7):e40390.
- 1846 Zhou, H., Friedman, H. S., and Heydt, R. v. d. (2000). Coding of border ownership in  
1847 monkey visual cortex. *The Journal of Neuroscience*, 20(17):6594–6611.