Solving the Binding Problem in Vision

Introduction

I propose to develop neural network computer simulations to explore potential solutions to how the primate visual system solves 'the binding problem'. I will also perform complementary visual psychophysical studies with human subjects to test the computational theories embodied in the computer models.

The binding problem is expressed by different authors in rather different ways. However, it generally boils down to the question of how the visual system represents which features are bound together as part of the same object. For example, if the two letters T and L are seen together, how does the visual system represent which horizontal and vertical bars are part of which letter? The Oxford Centre for Theoretical Neuroscience and Artificial Intelligence (OCTNAI) has developed a hierarchical neural network model, VisNet, of the primate ventral visual pathway [1-8]. Currently, this network model would represent the horizontal and vertical bars in the lower visual layers, and the letters in the higher visual layers. But so far there is no way of reading off which bars are part of which letters. This kind of feature binding is still an unsolved problem. Moreover, the binding of visual features must operate at all scales within a visual scene. How features are bound together must underpin how we segment a visual scene into objects and parts of objects, and thus how we make sense of the visual world. Duncan & Humphreys (1989) provide a good description of this process as follows [9]:

"A fully hierarchical representation is created by repeating segmentation at different levels of scale [10, 11]. Each structural unit, contained by its own boundary, is further subdivided into parts by the major boundaries within it. Thus, a human body may be subdivided into head, torso, and limbs, and a hand into palm and fingers [10]. Such subdivision serves two purposes. The description of a structural unit at one level of scale (animal, letter, etc.) must depend heavily on the relations between the parts defined within it (as well as on properties such as color or movement that may be common to the parts). Then, at the next level down, each part becomes a new structural unit to be further described with its own properties, defined among other things by the relations between its own sub-parts. At the top of the hierarchy may be a structural unit corresponding to the whole input scene, described with a rough set of properties (e.g., division into light sky above and dark ground below)." - Duncan & Humphreys (1989).

So what might be missing in the current VisNet architecture that might permit the network to begin to represent the essential binding information needed to comprehend the relationships between visual features and objects within complex visual scenes? The current VisNet model has the following two outstanding deficiencies.

Firstly, the current architecture of VisNet incorporates only bottom-up synaptic connections between successive neuronal layers. However, the primate ventral visual pathway is known to contain extensive top-down connections between layers and lateral connections within layers. It has previously been suggested that the top-down connections might implement attention to objects during visual search [12]. However, we now propose that top-down connections might play a far more fundamental role in visual processing. Specifically, it is possible that top-down connections guide competitive learning in lower layers, thus driving the formation of lower level visual representations that are modulated by higher level representations. For example, a simple cell in area V1 representing a vertical edge might learn to be modulated by top-down connections from higher level representations of object shape in, say, area V4 [13, 14]. In this case, the simple cell might respond only when the edge is on either the left or the right of an object, even though the object extends far outside the classical receptive field of the simple cell. This is indeed what has been observed in neurophysiology studies [15], thus confirming top-down modulation of cell responses. This kind of modulation may endow such neurons with additional information about binding within a scene. Below we expand on how competitive learning may utilise top-down connections between layers and lateral connections within layers to drive the development of 'binding neurons' that carry information about the binding relationships between low-level features and high-level features at every spatial scale.

Secondly, the current VisNet model is rate-coded. This means that, in order to save computational expense, the model does not explicitly simulate the timings of the electrical pulses, called action potentials or 'spikes', that pass between

neurons in the brain. However, real neurons in the brain not only communicate by exchanging such spikes, but also the way in which synapses are strengthened and weakened during learning are actually dependent on the timings of the spikes emitted by the pre- and post-synaptic neurons [16, 17]. For example, in the brain, a synapse may be strengthened if the pre-synaptic spike occurs about 20ms before the post-synaptic spike, but weakened if the pre-synaptic spike occurs about 20ms after the post-synaptic spike. This is known as spike time dependent plasticity (STDP). Because of this, spiking neural networks can give rise to radically different self-organisation of visual representations when trained on visual scenes containing multiple objects [18, 19]. Furthermore, if distributions of axonal delays between neurons are incorporated, then this can give rise to a phenomenon known as 'polychronization' [20]. This phenomenon involves the network learning many memory patterns, each of which takes the form of a repeating temporal loop of neuronal firings. These temporal memory loops self-organise automatically when STDP is used to modify the strengths of synapses in a recurrently connected spiking network with randomised distributions of axonal conduction delays between neurons. Polychronization can dramatically increase the selectivity of neurons and increase the memory capacity of a network. Below we discuss how polychronization may help the network to learn transform (e.g. location and view) invariant representations of visual objects that are not susceptible to false stimuli created by rearranging the elemental object features, and how this mechanism may also contribute to the development of highly selective binding neurons.

I propose, for the first time, to explore the effects of extending the VisNet model to include top-down connections between layers and lateral connections within layers, and spiking dynamics with spike time dependent plasticity. I will carry out a staged research programme that begins by examining the role of temporal polychronization in learning transform invariant object recognition, extend this to investigating the role of polychronization in the development of binding neurons, examine how these mechanisms explain experimental findings on visual search tasks [9], and then investigate how the proposed binding mechanisms may help to represent the feature relationships within faces, 3-dimensional objects, and 3-dimensional spatial environments. I will also carry out a series of visual psychophysical studies with human subjects to test the theories embodied in the computer models. There are a number of empirical avenues to investigate, which include exploring how the efficiency of visual search depends on the similarity relations between the targets and nontargets [9], the effect of pre-learning the binding relations of target objects on visual search [21], and the effects of localised brain lesioning on visual search to shed light on the separate contributions that different areas may make to visual binding [22, 23].

Proposed Research Programme

(1) Learning transform (e.g. location) invariant representations of structural descriptions of visual objects by temporal polychronization

OCTNAI has used VisNet to investigate two biologically plausible learning mechanisms that are able to train neurons to learn to respond to objects and faces with transform (e.g. location or view) invariance. The first of these learning mechanisms is known as 'trace learning' [1, 24]. This mechanism uses a temporal trace learning rule, which encourages neurons to learn to respond to images that occur close together in time. If, under natural viewing conditions, different transforms of an object are seen in temporal proximity, then trace learning forces individual neurons to respond to a particular object across all of its transforms. The second learning mechanism is 'continuous transformation (CT) learning' [3, 25]. This mechanism uses a standard Hebbian learning rule, which encourages neurons to learn to respond to images which are similar to each other. If the transforms of an object change continuously, then CT learning will also force individual neurons to respond to a particular object across all of its transforms. The implementation of both trace learning and CT learning in a spiking neural network has previously been explored by Evans & Stringer (2012) [18].

However, an outstanding problem with both the trace learning and CT learning mechanisms currently implemented in VisNet is that, after training, an output neuron that has learned to respond to a particular stimulus pattern with, say, translation invariance may also respond to false stimuli comprised of spatial rearrangements of the stimulus features. For example, if an output neuron is trained to respond to the letter T over a square region of the retina, then the output neuron will have developed strong afferent connections from V1 input neurons representing horizontal bars in all locations as well as V1 input neurons representing vertical bars in all positions. This means that the output neuron will be driven not only by the letter T, but also by any spatial rearrangement of the horizontal and vertical bars within the square training region. How might the output neuron learn to respond selectively just to the letter T across all locations

and not to false stimuli? This increased selectivity would likely also increase the stimulus capacity of the overall network.

I shall explore whether this might be achieved by temporal polychronization [20]. This can be effected by introducing distributions of axonal delays within the bottom-up, top-down and lateral connections, with the connection strengths self-organised using STDP. Such a network architecture is able to learn temporal attractor loops, in which the neurons comprising a memory pattern tend to fire in regularly repeating sequences. In the context of invariance learning, each temporal attractor loop would include the input neurons representing one transform (e.g. retinal location) of the visual object, the subset of output neurons that are learning to respond invariantly to that object, and any other hidden neurons that happen to be captured into the attractor state. Thus, during training each different transform (e.g. location) of the object would develop its own associated temporal attractor loop operating across the layers.

Polychronization dramatically increases the capacity of such an attractor network in terms of the number separate patterns that may be learned, where the same neuron may reoccur in many patterns without interference between these attractor states [20]. We therefore hypothesise that if the network is presented with a false stimulus constructed from an arbitrary spatial rearrangement of the object's features, then this will fail to drive one of the previously learned temporal attractor loops. In this case, polychronization will enable output neurons to learn to respond selectively to all of the transforms of a particular object, but not respond to false stimuli constructed through spatial rearrangement of the object's features. This improved selectivity should increase the object capacity of the network. I shall explore the potential of polychronization to prevent neuronal responses to false stimuli developing during training with either trace learning or CT learning in a spiking network.

(2) A new hypothesis for how the visual system solves the binding problem: the role of 'binding neurons'

Another limitation of the current purely bottom-up VisNet architecture is that the firing rates of the neurons are not able to specify which features are part of which objects. For example, consider again an image consisting of the two letters L and T. On the V1 input layer, each of these two letters is represented by one horizontal bar and one vertical bar. So when the two letters are presented together, there will be separate representations of two horizontal bars and two vertical bars. In the output layer, there would be separate representations of the letters L and T. However, it is not possible to read off from the neuronal firing rates within the network which horizontal and vertical bars are part of which letters. The correct binding between the bars and the letters could be determined if it was possible for neurons in the higher layers of the network to interrogate not only which bar and letter representations were active, but also which synapses between them were active thus specifying which bar representations were driving which letter representations. However, it is not possible for higher level neurons in the network to read off the activity of synapses in such a direct manner.

I shall explore the following hypothesis. Assume that, in addition to bottom-up connections between successive layers, the network also incorporates additional top-down connections between layers and lateral connections within layers. So the architecture is now more like a big attractor network. Also assume that the neurons in the network are spiking, with random distributions of axonal delays between neurons, and with synaptic connections modified by spike time dependent plasticity. These are the architectural requirements for temporal polychronization [20]. In this case, it is hypothesised that during visually-guided competitive learning, some random subset of neurons in every layer of the network, which we shall call 'binding neurons', will become tuned to respond when a particular low-level feature such as a bar is driving the representation of a specific high-level feature or object such as a letter T or L. Each different binding neuron will represent the binding relationship between a different pair of low-level feature and high-level feature or object.

It is important that the binding neuron responds if and only if the neurons representing the low-level feature are actually participating in driving the neurons representing the high-level feature. The binding neuron should not respond if the neurons representing the low-level feature and the neurons representing the high-level feature just happen to be co-active, where the former are not actually driving the latter. Such unrelated co-activation of low and high-level features might occur, for example, because of the presence of multiple similar objects within a complex natural scene. For example, if a T and L are presented together, then the neurons representing the horizontal bar of the T are co-active with the neurons representing the letter L, but the former are not driving the latter. So the corresponding binding neuron, which would represent that the given horizontal bar was part of the L, should not fire. This response specificity of the binding neuron is hypothesised to require temporal polychronization.

After training, the presentation of a letter T to the network would involve activating the input representations of the two bars comprising that letter. The two bars then drive the letter representation to respond in the higher layers. This will then drive the corresponding binding neurons to respond, where each binding neuron signals that a particular one of the bars is part of that letter. Moreover, if the representation of the letter is transform (e.g. location) invariant, then a particular constituent bar may be in different retinal locations depending on the location of the letter. In this case, different binding neurons could be tuned to different retinal locations of the bar, thus carrying spatial information about the location of the bar and hence the letter. This process could occur through polychronization within a spiking network by activation of a temporal attractor loop for each type of binding neuron similar to that described above. The temporal loop would incorporate the bar representation, the letter representation, and the corresponding binding neurons. I shall begin to explore this hypothesis by running simulations which investigate the binding of visual features comprising the alphabetic letters T and L.

With the inclusion of bottom-up, top-down and lateral connections, there are a variety of possible local network architectures that could self-organise through competitive learning to implement this, with the binding neurons being in any of the nearby lower or higher layers. Wherever the binding neuron is, its activation would still represent that a particular low-level feature is driving the representation of a specific high-level feature or object, and is therefore part of the object. A population of such binding neurons would specify which low-level features within a scene were part of which high-level features or objects, and this information could be read out directly by higher level neurons in the network. This process could operate at every spatial scale in the visual scene and at every level of the feature hierarchy within the network. A rich tapestry of binding neurons through the layers could help to provide a rich hierarchical structural description of a scene, perhaps rather analogous to that described above by Duncan & Humphreys (1989). This proposal may explain why the visual system needs extensive top-down connections between layers and lateral connections within layers, in addition to the bottom-up connections so far included in the current VisNet model.

Experimental evidence for this proposal has come from single unit recording of V1 simple cells, which appear to be modulated by image context outside of their classical receptive field. It has been found that the responses of some V1 simple cells are modulated by which side of an object boundary they represent, with some V1 cells responding most if they represent the left boundary and other V1 cells responding most if they represent the right boundary [15]. In these cases, it is possible that the responses of these V1 simple cells are modulated by top-down connections from a higher layer of the ventral visual pathway such as V4. Area V4 contains neurons that represent local boundary elements of objects, and the responses of these neurons are sensitive to where the boundary element is with respect to the centre of mass of the object [13, 14]. So top-down connections from these V4 cells could explain the modulation of V1 simple cells by which side of an object they are representing. Such modulated V1 simple cells could be considered as examples of one kind of binding neuron, representing that a particular edge is on the left or right of an object.

The discussion above suggests that binding is a much richer phenomenon than traditionally described by visual psychologists. Indeed, the binding mechanism proposed here is potentially so rich that it would be impossible to describe the process at a high psychological level; it requires a description at the neuronal level. Furthermore, if the semantic analysis of visual scenes carried out by the primate brain utilises these hypothesised binding neurons, then such neurons could not develop in biologically implausible neural network architectures such as backpropagation of error networks. Although backpropagation networks might be efficient at learning arbitrary mappings, and solving narrow tasks like face recognition, they would not be able to represent the essential binding information needed to semantically analyse visual scenes in the same way as the primate brain.

(3) The roles of binding and temporal polychronization in visual search

Some researchers have tried to relate feature binding to the speed of visual search for target objects among nontarget distractors. For example, Treisman & Galade (1980) proposed Feature Integration Theory [26], which assumes that there is only a single locus of attention where visual features are bound together. This would require a serial search for a visual search task that requires feature binding, but allows faster parallel search for other search tasks that do not require feature binding. In contrast, we have proposed that binding is carried out by binding neurons that operate simultaneously across the whole visual field. In this case, there is no need for binding to be limited to a single spatial locus of attention, and the time taken for visual search would never be governed by the need to perform a serial search

with a single locus of attention. Instead, binding may operate in parallel across the visual field, and the search time would be related to other factors determining the intrinsic difficulty of the task. This is supported by the study of Duncan & Humphreys (1989) [9]. These authors found no clear dichotomy between serial and parallel modes of search. This contradicts the assumption of Feature Integration Theory that there is a single locus of feature binding, which leads to serial search for those tasks that require feature binding and parallel search for tasks that do not. Instead, Duncan & Humphreys (1989) reported that search efficiency was found to decrease as the targets became more similar to nontargets, or if the nontargets became more dissimilar to each other. Duncan & Humphreys (1989) suggested that the neural basis of these findings is that "A parallel stage of perceptual grouping and description is followed by competitive interaction between inputs, guiding selective access to awareness and action".

I propose to model the visual search results of Duncan & Humphreys (1989). In our model, multiple objects within a visual scene, say a target and a number of nontargets, would be represented by separate hierarchies of visual features through the layers of the network, where the low-level features are bound to the high-level features by activity loops involving additional binding neurons. This would provide a rich representation of natural scenes with all of the required binding information. As suggested by Duncan & Humphreys (1989), these multiple object representations would compete with each other through inhibitory interneurons within each layer of the hierarchical network. However, the competition between objects would also be affected by top-down attention to the target object. Specifically, in order to model visual search for the target object, we will apply a top-down excitatory signal to neurons representing this object in the higher layers of the network. This would preferentially facilitate the hierarchical representation of the target object through the layers via the top-down connections. The facilitated representation of the target object could extend down to the earliest layers of the network, thus providing spatial information about target location (e.g. for a following saccade).

Now let us consider the two experimental findings from Duncan & Humphreys (1989). As suggested by these authors, we may be able to explain their findings by considering the competitive interactions between the object representations. First, the difficulty of visual search may increase with increased similarity of targets to nontargets because the top-down attentional signal will drive many features associated with the nontarget objects. This could either lengthen the time it takes for the network to settle into a state where the target object dominates, or might lead to a final state where there is more residual activity for the nontarget objects compared to when the nontarget objects are dissimilar to the target object. Secondly, difficulty may increase with decreased similarity between nontargets because as the nontargets become more dissimilar they will activate many more different kinds of visual feature through the layers, thus increasing the general level of competition with layers. This will inhibit the representation of the target object.

Lastly, although our theory permits feature binding to operate in parallel across the entire visual field, it would still be expected that visual processing, which includes feature binding, would be somewhat degraded away from the spatial locus of attention. This could occur because the neural representation of the part of the visual scene at the site of spatial attention, which might be highlighted due to high acuity foveal fixation or top-down attentional facilitation, would compete strongly with visual processing of the rest of the scene through inhibitory interneurons. This strong inhibition from the attended part of the scene would likely degrade visual processing elsewhere, including feature binding operations. This would explain various psychophysical findings about binding in human vision [27]. However, this is quite different from the underlying assumption of Feature Integration Theory, which actually requires only a single spatial locus of attention to perform feature binding, and so cannot permit any binding elsewhere.

(4) Binding of visual features comprising faces

I will investigate how neurons learn to respond selectively to faces without also responding to false stimuli constructed by spatial jumbling of the facial features, and how a set of individual facial features such as the eyes, nose and mouth are bound together to represent a whole face.

Neurophysiology studies carried out by Freiwald and colleagues [28] have begun to show how neurons encode individual facial features, such as the eyes, nose and mouth, as well as the spatial relations between these facial features. OCTNAI has begun to model how these neuronal responses may develop through visually-guided learning as

the network is trained on lots of faces. However, the simulations were carried out in the current version of VisNet, which is rate-coded with a purely feedforward (bottom-up) network architecture. I now propose to extend this work using a spiking neural network with bottom-up, top-down and lateral connections, and explore how this more realistic architecture may enhance face processing.

First, I will investigate how neurons may learn to respond to faces but avoid also responding to false stimuli constructed from arbitrary spatial rearrangements of the facial features. So far, OCTNAI has been using trace learning to develop translation invariant neuronal responses across a small number of distinct retinal locations. We will explore how the network performs on the more challenging situation in which it is trained on faces presented across a large continuum of retinal space. This may cause face cells in higher layers to form strong afferent connections from neurons in lower layers representing facial features across all retinal locations, and so increase the chance that face cells also respond inappropriately to arbitrary spatial jumbling of facial features. We will explore this in both a rate-coded network with Hebbian learning and a spiking network with STDP. If the rate-coded network fails to produce neurons that respond selectively to faces without responding to jumbled facial features, then we will explore whether the spiking network is able to solve this with the potentially enhanced neuronal selectivity that might be afforded by polychronization. In particular, we will investigate whether polychronization of neuronal assemblies across the layers helps to eliminate neuronal responses to false stimuli comprised of arbitrary spatial arrangements of facial features.

Second, I will investigate how facial features such as the eyes, nose and mouth may be bound to a particular face through the development of binding neurons, especially when the given face is one of multiple faces within a crowd scene. Another limitation of previous OCTNAI research is that VisNet was trained with only one face at a time. I will begin by investigating whether the same desired neuronal responses develop if the network is trained more realistically on multiple faces presented simultaneously. Do the faces have to shift independently on the retina for the desired neuronal responses to develop [4, 7, 29]? If the network is tested with multiple faces presented simultaneously, how can the network bind the individual facial features represented in the higher layers to the correct face? Are the facial features for each face synchronised with each other, with desynchronisation between faces [19]? Or is binding achieved by additional binding neurons encoding which low level features drive which high level features operating through the layers, as hypothesised above? I will therefore explore whether such binding neurons develop during training.

(5) Binding of visual features comprising 3-dimensional objects

Recent neurophysiology studies have revealed how particular populations of neurons within the primate ventral visual pathway represent the shapes of 2 and 3-dimensional objects. For example, some neurons in visual area V4 have been found to encode the boundary shapes of 2-dimensional objects, whereby individual neurons respond to local boundary elements of a particular curvature and position with respect to the centre of the object [13, 14]. The entire shape of the object may then be represented by a distributed subpopulation of these neurons, each encoding the conformation of a localised region of the boundary. Other kinds of neuron at a later stage of processing in the ventral pathway have been found to encode the boundary shapes of 3-dimensional objects, whereby individual neurons respond to local surface patches of a particular curvature and position with respect to the centre of the object [30]. The entire 3-dimensional shape of an object may then be represented by a distributed subpopulation of these neurons.

OCTNAI has developed the first biologically plausible neural network model of how the ventral visual pathway may learn to encode the boundary shapes of 2-dimensional objects, with individual neurons responding to local boundary elements. This was achieved by training the neural network model on many different 2-dimensional object shapes. As the number of objects increases, there is a statistical decoupling [4, 29] between different boundary elements defined by their curvature and position with respect to the centre of the object. This forces individual neurons in the higher layers to learn to respond to just one or two boundary elements, as observed in neurophysiology studies. We now propose to extend this work by introducing stereoscopic vision to the neural network model, and then training the model on many 3-dimensional objects. We hypothesise that a similar learning mechanism driven by statistical decoupling will force individual neurons in the higher layers to learn to respond to one or two surface patches. I will

carry out a detailed examination of the varied spatial forms of the surface patches encoded by neurons in the model, and how they may provide a distributed encoding of arbitrary 3-dimensional object shapes.

How might the primate visual system represent multiple distinct 3-dimensional objects presented together simultaneously? Consider the situation where two objects, a chair and table, are seen together. The shape of each object is represented by a separate distributed subpopulation of neurons, where each neuron encodes a local surface patch of one of the objects. How does the network represent which surface patches are part of which object? This is an example of a binding problem. As described above, we propose that this may be solved by the presence of additional neurons, termed 'binding neurons', which only respond when neurons representing a specific surface patch are driving neurons representing a particular object, which implies that the surface patch is part of the object. These binding neurons will carry measurable information about the fact that the surface patch is part of the object. I will explore the development of these binding neurons during visually-guided training, and the kind of binding information that these neurons carry about the 3-dimensional spatial structure of objects. We conjecture that the model may learn to represent each object using a deep hierarchy of 3-dimensional spatial features, with high-level (larger) spatial features comprised of multiple low-level (smaller) spatial features, and where the binding relationships between these features are represented by binding neurons.

(6) Binding of visual features comprising 3-dimensional spatial environments

We hypothesise that the learning mechanisms described above may be extended to investigating how the primate brain learns to represent the egocentric 3-dimensional spatial structure of the global visual environment. In this study, instead of training the network model on single objects, the network will be trained on many different 3-dimensional spatial environments each comprised of multiple surrounding objects of various shapes. Across many different environments, there will be a statistical decoupling between the elemental 3-dimensional spatial features, such as flat surfaces, edges and corners, which comprise those environments. According to our theory, individual neurons in the higher layers of the network should learn to encode one or two of these elemental spatial features. I will carry out a detailed analysis of the varied forms of the spatial features encoded by neurons in the model, and how they may provide a distributed egocentric encoding of arbitrary 3-dimensional environments.

I will also investigate how the network model might solve the associated binding problem. That is, how might the network represent which lower-level features, such as a corner, are part of which higher-level features, such as a table? We hypothesise that if the network contains a mixture of feedforward, feedback and lateral connections between neurons in successive layers, then during training some random neurons, which we term binding neurons, will learn to respond only when neurons representing a specific lower-level feature are driving neurons representing a particular higher-level feature, which implies that the lower-level feature is part of a higher-level feature. These binding neurons will carry measurable information about the fact that the lower-level feature is part of the higher-level feature. We will explore the development of these binding neurons during visually-guided training, and the kind of binding information that these neurons carry about the 3-dimensional spatial structure of the environment. Similar to above, we conjecture that the model may learn to represent each major object or large scale feature within the environment using a deep hierarchy of 3-dimensional spatial features, with high-level (larger) spatial features comprised of multiple low-level (smaller) spatial features, and where the binding relationships between these features are represented by binding neurons.

(7) Psychophysical studies with human test subjects

I will also carry out a series of visual psychophysical studies with human subjects to test the theories embodied in the computer models. There are a number of empirical avenues to investigate.

(i) I shall explore the performance of human test subjects on a range of visual search tasks in which the similarity relations between the targets and nontargets are systematically varied, and compare these experimental findings with the results of computer simulations. The human psychophysical studies of Duncan & Humphreys (1989) [9] found no clear dichotomy between serial and parallel modes of visual

search. Instead, search efficiency was found to decrease as the targets became more similar to nontargets, or if the nontargets became more dissimilar to each other. I shall extend this work by examining the effects of the similarity relations between target and nontarget stimuli, and between nontarget and nontarget stimuli, on visual search tasks performed by human subjects for a range of both 2D and subsequently 3D properties of visual stimuli. These experimental findings will be compared with the results of complementary computer simulations.

(ii) I shall explore the effects of pre-learning the binding relations of target objects on visual search tasks carried out by human test subjects, and compare these experiments with the results of computer simulations. Feature Integration Theory [26] suggests that different kinds of low-level visual features, such as stimulus orientation and colour, are initially computed independently. The subsequent step of binding features to generate a coherent perception of a stimulus is generally thought to be rate limited by the need to process one spatial location at a time. Our own theory of parallel binding via 'binding neurons' across the visual field also recognises that the most efficient binding may in practice still be localised to a single spatial locus of attention due to the effects of competitive interactions. Rappaport et al (2013) [21] examined whether these processing limitations remain once bindings are pre-learned for familiar target objects. Specifically, they looked at the effects of stored knowledge about colour-shape relations, and how these facilitate binding during visual search. Participants searched for objects that could appear either in familiar or unfamiliar colours. If the target objects were presented in their pre-learned colours, then they were detected at rates consistent with efficient parallel binding across multiple stimuli. Presenting target objects in unfamiliar colours led to less efficient search. These findings suggest that the visual system was able to exploit the pre-learned representations of feature conjunctions, and thereby reduce attentional constraints during visual search. At first sight, these results seem potentially consistent with both Feature Integration Theory (extended to allow pre-learning of familiar conjunctions of features to form new independent feature conjunction maps) and our own theory of simultaneous binding across the visual field via binding neurons. I shall extend these psychophysical studies on human subjects, in which the feature bindings are pre-learned for target objects, and compare the experimental findings with the results of complementary computer simulations. Can we find evidence to differentiate between the two theories? The binding neuron theory hypothesises that much more detailed binding information will be available to the human subject across a visual scene, such as the full hierarchical structural descriptions and retinal locations of multiple stimuli throughout the visual field.

(iii) I shall explore the effects of brain lesioning on the time course of visual search in human subjects in order to shed light on the separate contributions that different brain areas may make to different stages and kinds of visual binding, and compare these findings with the results of complementary computer simulations. Humphreys et al. (2009) [22] carried out visual search studies with control participants and patients with unilateral parietal or fronto/temporal lesions. They found that parietal patients performed like the other participants at detecting targets involving within-domain conjunctions of form, but were impaired at detecting targets involving cross-domain conjunctions of form, colour and size. These studies provided evidence that the cross-domain binding of form, colour and size is a distinct process, which is carried out in separate brain areas from the within-domain binding of form. In another psychophysical study, Braet and Humphreys (2009) [23] investigated the occurrence of illusory conjunctions (ICs) in a colour/letter identification task from a patient with bilateral parietal lesions, and from a transcranial magnetic stimulation (TMS) study with normal participants. Their data suggested that the posterior parietal cortex plays a key role at a late stage of feature binding. I shall extend these lesion studies on human subjects in order to provide a more fine-grained analysis of how different aspects of visual binding are performed in separate stages in different brain areas, and compare the experimental findings with the results of complementary computer simulations.

Summary

I propose to explore a novel computational solution to how the primate visual system may solve the binding problem. This is one of the outstanding grand challenges of theoretical neuroscience. The hypothesis of 'binding neurons' discussed above is biologically plausible in that the mechanism relies on a biologically plausible network architecture with local STDP synaptic learning rules. I plan to investigate how this binding mechanism operates on a number of test problems, including the representation of alphabetic letters, faces, 3-dimensional objects and 3-dimensional environments. A particularly interesting feature of the proposed theory is that it potentially reveals a sharp contrast between brain processing and the operation of biologically implausible neural network algorithms such as

backpropagation of error. The architecture of the brain, and consequently our own brain-inspired neural network models, will support the development of 'binding neurons' that represent the binding relationships between low-level and high-level features at all spatial scales throughout a visual scene. However, a biologically implausible neural network algorithm such as backpropagation of error cannot develop binding neurons and so cannot represent any such binding information. Instead, backpropagation networks simply learn a fixed mapping from inputs to outputs, for example, to determine whether a face is present within a visual scene. Consequently, these kinds of engineering approaches will ultimately be unable to process complex natural visual scenes with the same semantic richness as the human brain. Future generations of researchers working in computer vision and robot control will need to look to the architecture and workings of the brain to build lifelike machines that are able to comprehend their world and act flexibly within it. My proposed research on visual binding is set to demonstrate and establish this fact for everyone interested in the future of artificial intelligence.

References

1. Wallis, G. and E.T. Rolls, *Invariant face and object recognition in the visual system.* Prog Neurobiol, 1997. **51**(2): p. 167-94.
2. Stringer, S.M. and E.T. Rolls, *Position invariant recognition in the visual system with cluttered environments.* Neural Netw, 2000. **13**(3): p. 305-15.
3. Stringer, S.M., et al., *Learning invariant object recognition in the visual system with continuous transformations.* Biol Cybern, 2006. **94**(2): p. 128-42.
4. Stringer, S.M. and E.T. Rolls, *Learning transform invariant object recognition in the visual system with multiple stimuli present during training.* Neural Netw, 2008. **21**(7): p. 888-903.
5. Tromans, J.M., M. Harris, and S.M. Stringer, *A computational model of the development of separate representations of facial identity and expression in the primate visual system.* PLoS One, 2011. **6**(10): p. e25616.
6. Tromans, J.M., I. Higgins, and S.M. Stringer, *Learning view invariant recognition with partially occluded objects.* Front Comput Neurosci, 2012. **6**: p. 48.
7. Tromans, J.M., H.J. Page, and S.M. Stringer, *Learning separate visual representations of independently rotating objects.* Network, 2012. **23**(1-2): p. 1-23.
8. Higgins, I.V. and S.M. Stringer, *The role of independent motion in object segmentation in the ventral visual stream: Learning to recognise the separate parts of the body.* Vision Res, 2011. **51**(6): p. 553-62.
9. Duncan, J. and G.W. Humphreys, *Visual search and stimulus similarity.* Psychol Rev, 1989. **96**(3): p. 433-58.
10. Marr, D. and H.K. Nishihara, *Representation and recognition of the spatial organization of three-dimensional shapes.* Proc R Soc Lond B Biol Sci, 1978. **200**(1140): p. 269-94.
11. Palmer, S.E., *Hierarchical structure in perceptual representation.* Cognitive Psychology, 1977. **9**(4): p. 441-474.
12. Deco, G. and T.S. Lee, *A unified model of spatial and object attention based on inter-cortical biased competition.* Neurocomputing, 2002. **44–46**(0): p. 775-781.
13. Pasupathy, A. and C.E. Connor, *Shape representation in area V4: position-specific tuning for boundary conformation.* J Neurophysiol, 2001. **86**(5): p. 2505-19.
14. Pasupathy, A. and C.E. Connor, *Population coding of shape in area V4.* Nat Neurosci, 2002. **5**(12): p. 1332-8.
15. Zhou, H., H.S. Friedman, and R. von der Heydt, *Coding of border ownership in monkey visual cortex.* J Neurosci, 2000. **20**(17): p. 6594-611.
16. Bi, G.Q. and M.M. Poo, *Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type.* J Neurosci, 1998. **18**(24): p. 10464-72.
17. Markram, H., et al., *Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs.* Science, 1997. **275**(5297): p. 213-5.
18. Evans, B.D. and S.M. Stringer, *Transformation-invariant visual representations in self-organizing spiking neural networks.* Front Comput Neurosci, 2012. **6**: p. 46.
19. Evans, B.D. and S.M. Stringer, *How lateral connections and spiking dynamics may separate multiple objects moving together.* PLoS One, 2013. **8**(8): p. e69952.
20. Izhikevich, E.M., *Polychronization: computation with spikes.* Neural Comput, 2006. **18**(2): p. 245-82.

21. Rappaport, S.J., G.W. Humphreys, and M.J. Riddoch, *The attraction of yellow corn: reduced attentional constraints on coding learned conjunctive relations.* J Exp Psychol Hum Percept Perform, 2013. **39**(4): p. 1016-31.
22. Humphreys, G.W., J. Hodsoll, and M.J. Riddoch, *Fractionating the binding process: neuropsychological evidence from reversed search efficiencies.* J Exp Psychol Hum Percept Perform, 2009. **35**(3): p. 627-47.
23. Braet, W. and G.W. Humphreys, *The role of reentrant processes in feature binding: Evidence from neuropsychology and TMS on late onset illusory conjunctions.* Visual Cognition, 2009. **17**(1-2): p. 25-47.
24. Földiák, P., *Learning Invariance from Transformation Sequences.* Neural Computation, 1991. **3**(2): p. 194-200.
25. Perry, G., E.T. Rolls, and S.M. Stringer, *Continuous transformation learning of translation invariant representations.* Exp Brain Res, 2010. **204**(2): p. 255-70.
26. Treisman, A.M. and G. Gelade, *A feature-integration theory of attention.* Cogn Psychol, 1980. **12**(1): p. 97-136.
27. Wolfe, J.M. and K.R. Cave, *The psychophysical evidence for a binding problem in human vision.* Neuron, 1999. **24**(1): p. 11-7, 111-25.
28. Freiwald, W.A., D.Y. Tsao, and M.S. Livingstone, *A face feature space in the macaque temporal lobe.* Nat Neurosci, 2009. **12**(9): p. 1187-96.
29. Stringer, S.M., E.T. Rolls, and J.M. Tromans, *Invariant object recognition with trace learning and multiple stimuli present during training.* Network, 2007. **18**(2): p. 161-87.
30. Yamane, Y., et al., *A neural code for three-dimensional object shape in macaque inferotemporal cortex.* Nat Neurosci, 2008. **11**(11): p. 1352-60.