Perception and action without 3D coordinate frames

Andrew Glennerster
School of Psychology and Clinical Language Sciences, University of Reading, Reading RG6 6AL, UK
a.glennerster@reading.ac.uk
http://www.personal.reading.ac.uk/~sxs05ag/

James Stazicker
Department of Philosophy, University of Reading, Reading RG6 6AA, UK
j.stazicker@reading.ac.uk
https://jamesstazicker.com/

Abstract

Neuroscientists commonly assume that the brain generates representations of a scene in various non-retinotopic 3D coordinate frames, for example in 'egocentric' and 'allocentric' frames. Although neurons in early visual cortex might be described as representing a scene in an eye-centred frame, using 2 dimensions of visual direction and one of binocular disparity, there is no convincing evidence of similarly organized cortical areas using non-retinotopic 3D coordinate frames nor of any systematic transfer of information from one frame to another. We propose that perception and action in a 3D world could be achieved without generating ego- or allocentric 3D coordinate frames. Instead, we suggest that the fundamental operation the brain carries out is to compare a long state vector with a matrix of weights (essentially, a long look-up table) to choose an output (often, but not necessarily, a motor output). The processes involved in perception of a 3D scene and action within it depend, we suggest, on successive iterations of this basic operation. Advantages of this proposal include the fact that it relies on computationally well-defined operations corresponding to well-established neural processes. Also, we argue that from a philosophical perspective it is at least as plausible as theories postulating 3D coordinate frames. Finally, we suggest a variety of experiments that would falsify our claim.

Key words: 3D representation; coordinate frames; moving observer; stereopsis; motion parallax.

1. Introduction

For an observer to perceive and operate successfully in a 3D world it is often assumed that the brain must generate 3D models or reconstructions of the scene in a range of coordinate frames (retinotopic, head-centred, hand-centred, body-centred, world-centred) and that there must be a series of complex 3D coordinate transformations to convert information from one frame to another, even before the sensory information is passed to the motor system (Andersen et al, 1997; Colby, 1998; Byrne et al, 2007). The hypothesis explored in this article is that these coordinate transformations may be avoided in the brain, and that an alternative account could explain 3D perception of surface shape, object location and scene layout.

We focus particularly on the problem of specifying the layout of a 'vista space' (Meilinger, 2008), i.e. the scene that is visible from one place as an observer moves their head and eyes around freely.

We argue that the brain solves this problem without exploiting either an egocentric or a world-based 3D coordinate frame representation of the space. Instead, we propose, neurophysiological mechanisms of the type that are already recognised as underpinning the control of movement are critical in explaining perception of the 3D environment. These mechanisms were originally proposed in the 1960s to explain the operation of the cerebellum (Marr 1969; Albus 1971, Schmahmann, 2004). They describe how a sensory state can be 'recognized' (i.e. identified as most similar to a stored sensory context) and hence give rise to a motor output associated with that stored context.

Similar mechanisms to the one we propose have been suggested for the representation of space over a larger range of movements (Gillner and Mallot 1998; Franz et al. 1998; Cummins and Newman 2008; Milford and Wyeth 2008) and our broader aim is to provide a common framework to understand spatial perception over a wide range of scales without resorting to 3D coordinate frames at any stage. However, in this article we focus on the perception of vista space which requires (i) saccadic eye movements or other rotational movements to allow the observer to see the entire scene around them (optic array) and (ii) relatively small translations of the optic centre while fixating an object, which includes lateral head movements and static binocular viewing of an object. In a static world, these movements, along with hand movements and larger translations that change the optic array completely, cover a large proportion of image changes that an observer can generate. All of these movements can be described in terms of a graph of views, where the nodes are views and the edges of the graph are actions, an idea that is well established in models of navigation (Franz et al 1998). We do not restrict the discussion to vision. Instead, all sensory modalities contribute to a 'sensory context' which is quite different from bringing all modalities into a common 3D coordinate frame or reference frame as others have suggested (Jay and Sparks 1987; Andersen et al 1997; Cohen and Andersen 2002).

A second novel aspect of the proposal is that we suggest 'motivational' input is included as part of a description of the current state (hence we talk of a 'sensory+motivational' state) and the same is true of the stored contexts. The term 'motivational' is intended to distinguish this part of the input from input with an immediately sensory source. Motivational input includes information about the observer's task, but it need not be limited to the observer's drives or desires.

We will set out three reasons for accepting our proposal. First and foremost, the mechanism we propose can provide an account for performance in tasks in vista space that are standardly explained by appeal to representation in 3D coordinate frames. This includes representation of the slant, depth relief and relative distance of surfaces by a moving observer (see *A 2½D sketch* below). Second, while it is an unresolved question as to how the brain could implement the transformations required by a 3D coordinate frame theory, it is well established that neurons can implement the type of process that is invoked in our proposal (see Section 5). Third, participants' performance in spatial judgement tasks suggests that their representations of vista space cannot be captured in terms of a single, consistent 3D coordinate frame representation of that space (Section 2). We will contrast our approach with other proposals in neuroscience, giving examples of physiological and psychophysical findings that could discriminate between the rival hypotheses. We will also relate our proposal to philosophical theories of perceptual experience, and consider some philosophical and neuroscientific challenges (Section 6).

In the next section, we introduce a distinction from the philosophical literature that enables us to specify the difference between our proposal and one based on coordinate frames. We also explain our proposal's implications for philosophical theories of perceptual experience. Then, in Section 3, we provide a notation to define our proposal more formally.

2. 3D representation: 3D coordinate frame or not?

Our proposal is that 3D perception and action in a vista space are achieved through a process of matching the current sensory+motivational context with stored sensory+motivational contexts, and that in order for perception and action to be achieved in this way, no 3D coordinate frame representation of the vista space is required. It is worth emphasizing that both conjuncts of this proposal are essential to it. By itself, the claim that perception and action are achieved through a process of matching current and stored sensory+motivational contexts is extremely general. As a result, by itself this claim is consistent with representation in 3D coordinate frames; the contexts in question could *be* 3D coordinate frame representations. Our proposal is substantive and falsifiable, in that it involves both the positive claim that 3D perception and action in a vista space are achieved through the matching process, and the negative claim that this does not require representation in 3D coordinate frames.

So what is representation in a 3D coordinate frame? To make this precise, we need to distinguish between the *content* and the *format* of a representation. Roughly, a representation's content is the information that the representation makes available to a system. The content of a representation may be modeled as the possible state of the world that obtains wherever the representation is correct, for example the presence of a surface with a certain shape and size at a certain distance and direction from the observer.[1] By contrast, the format of a representation is the means by which the vehicle of representation carries its content. For example, a map and a sentence might each have the same content – they might tell you the same facts about how

[1] Some philosophers would also include a *mode of presentation* of a possible state of the world, a further aspect of the representation's content which contributes to its functional role. We try to keep things simple by working in terms of extensional contents (i.e., contents which may be modeled simply as possible states of the world).

objects are arranged, say – though they represent that content through different representational formats (Camp, 2007).

Our negative claim is that the contents of perceptual representations do not, in general, include a 3D coordinate frame representation of the whole of vista space. That is, the states of the world that perceptual processes represent do not, in general, include certain spatial relations: relations between objects (or their parts) on the one hand, and an origin and three axes on the other, extending throughout the vista space around the observer.[2] Individual objects and subregions of the vista space are perceived as having height, breadth and depth, dimensions which may be plotted against an origin and three axes. But we deny that, in general, perceptual processes represent the entire vista space in this way, with spatial relations among all locations in the space represented in a way that may be plotted against a single origin and three axes. We accept that, for some very specific tasks, representations of these relations throughout a vista space might be constructed. But we deny that perception and action in a 3D world in general rely on that construction. Our alternative can also be contrasted with 3D coordinate frame representation in that the latter, but not the former, is an essentially metric form of representation (see Section 4).

Though our negative claim concerns the contents of perceptual representation, it might be falsified by findings about the format of perceptual representation: if perceptual representations had a format like that of a 3D model of a vista space, this would be strong evidence that these representations exploit that format to carry contents that are expressible in terms of 3D coordinate frames (see Section 5). For example, one way to describe the receptive fields of disparity tuned neurons in V1 is that each cell can be mapped to a voxel in a coordinate frame defined by three axes, where the origin lies at the fixation point and variation along the horizontal and vertical axes

[2] If a coordinate frame system did represent the entire vista space in this way, it might be unspecific or silent about what is present at some coordinates in the vista space. So our negative claim is not supported merely by the fact that some regions of a vista space are occluded.

is determined by a cell's spatial location in the cortex while variation along the depth axis is determined by a cell's disparity sensitivity (Prince et al. 2002). This format in V1 does not undermine our hypothesis, since the V1 coordinate frame only applies to a subregion of vista space and the frame moves with the eye. But if a parallel format were found in brain areas that encode head-centred or world-centred frames, especially if it could be shown that these formats are inherited or computed on the basis of the V1 coordinate frame, that would be strong evidence against our hypothesis (see Section 6.3).

Philosophical work on spatial perception often assumes that the contents of perceptual representations include 3D coordinate frames (see e.g. the essays in Eilan, McCarthy and Brewer, 1993). Here is an influential example:

> I suggest that one basic form of representational content should be individuated by specifying which ways of filling out the space around the perceiver are consistent with the representation's content being correct. The idea is that the content involves a certain spatial *type.* … There are two steps we have to take if we are to specify fully one of these spatial types. The first step is to fix an origin and axes. … [F]or instance, one kind of origin is given by the property of being the center of the chest of the human body, with the three axes given by the directions back/front, left/right, and up/down with respect to that center. … Having fixed origin and axes, we need to take the second step in determining one of the spatial types, namely, that of specifying a way of filling out the space around the origin.
>
> Peacocke, 1992: 62-3

Peacocke suggests that the spatial contents of perception take the form of a 3D coordinate frame representation of the vista space around an observer. This is the kind of representation which we deny is required to explain perception and action in a vista space.

Like much of the philosophical literature, Peacocke appeals to perceptual representations in order to capture the phenomenology of perceptual experience; his core claim about 3D coordinate frame representation is that ‘the appropriate set of labeled axes captures distinctions in the phenomenology of experience’ (63). In what follows we do not focus on describing phenomenology. Instead we explain the perceptual achievements revealed by observers’ reports of the layout of a scene, in terms of a system of representations in the brain that does not include 3D coordinate frames. Nonetheless, our proposal undermines Peacocke’s claim that the phenomenology of perceptual experience should be characterized in terms of 3D coordinate frames. Observers’ reports about a scene are a key empirical measure of phenomenology or conscious experience (Weiskrantz 1997; Deheane and Changeux 2004), and our proposal is that the system of representations which best explains these reports is not a system of 3D coordinate frame representations.

Our explanation of participants’ reports is more powerful than an explanation in terms of 3D coordinate frames, because some reports which our proposal explains cannot be explained by any ‘way of filling out space’ relative to three axes and an origin. For example participants in Glennerster’s lab were presented with an expanding virtual scene, and made pairwise comparisons of the distances to objects. Although the pairwise comparisons were precise (i.e. highly repeatable), there was no consistent depth ordering of the objects that could explain this performance (e.g. A>B>D yet also A<C<D). No single 3D-coordinate plotting of the objects’ locations is consistent with this form of successful spatial perception (Svarverud et al., 2012), a conclusion others have proposed on the basis of similar apparently inconsistent psychophysical evidence (Koenderink et al, 2002; Smeets et al, 2009, see Section 5). Our explanation of participants’ reports is also simpler than an explanation in terms of 3D coordinate frames, because

in contrast with the transformations required by a 3D coordinate frame theory, the type of process we propose is one that we know neurons can implement (see Section 5).

These advantages turn on the positive part of our proposal, to which we turn next. Our positive claim is that perception and action are achieved through a process of matching the current sensory+motivational context with a very long list of stored sensory+motivational contexts. The next section specifies this more formally. The consequence of picking one context is an output (which usually results in a motor output) and hence a new context, a loop that is entirely familiar in the domain of motor control. We propose that this step is a fundamental one in the representation of 3D shape and layout, as we illustrate in Section 4.

3. Recognising a stored context

We consider every neural firing rate in the nervous system that might contribute to the next action as an element in a long list or vector, $\vec{r}$, with length n. This defines the current context. The nervous system stores synaptic weights in such a way that they can be compared to the firing rates listed in $\vec{r}$. The current context, $\vec{r}$, is compared to a large number of stored contexts, where each stored context is a long list of synaptic weights, also of length *n*. In Marr's model of the cerebellum (Marr, 1969), each of these stored contexts corresponded to the synaptic weights of a Purkinje cell but the idea of comparing an input vector of firing rates to a stored vector of synaptic weights is a general principle of neural computation that is applicable from the level of a single neurons up to very large sets of neurons (e.g. McCulloch and Pitts, 1943; Rolls and Treves, 1998). If there are *m* stored contexts then these synaptic weights form an *m* by *n* matrix, W, with each row constituting one stored context, $w_{(i,*)}$. For the sake of notational simplicity (and following Marr (1969)), we assume that the magnitude of $\vec{r}$ is the same as that of each stored context:

$$\| \vec{r} \| = \| w_{(i,*)} \| , \ \forall i = [1, \ldots, m] \qquad \text{Eq (1)}$$

For example, if each neuron contributing to $\vec{r}$ is either firing or not (1 or 0) and each synaptic weight is either 'on' or 'off' (1 or 0), this is equivalent to assuming that the proportion of neurons firing at any one time is constant (*k*) and equal to the proportion of synapses that are 'on' in each stored context, $w_{(i,*)}$. There is evidence that this proportion might be quite low (Attwell and Laughlin, 2001).

The process of comparing $\vec{r}$ (*n* by 1) to the stored contexts, W (*m* by *n*), is simply:

$$k = argmax_i(W\vec{r}) \qquad \text{Eq (2)}$$

where the function $argmax_i(\vec{x})$ returns the index of the maximum value in $\vec{x}$, i.e. k is the index to W that gives the maximum correlation between $\vec{r}$ and $w_{(i,*)}$ for any $i = [1, \ldots, m]$ and hence $w_{(k,*)}$ is the 'recognised context'. Geometrically, this amounts to searching on an n-dimensional sphere for the context, $w_{(i,*)}$, that is closest to $\vec{r}$ (in the sense of having smallest angle between them). Each context, $w_{(i,*)}$, is associated with a predetermined output (in the simple case, a motor output). In this sense, the list of stored contexts can be considered to be a list of 'reflexes'.

The proposal is neutral about the location in the nervous system of the relevant synaptic weights, W, and the firing rates, $\vec{r}$.

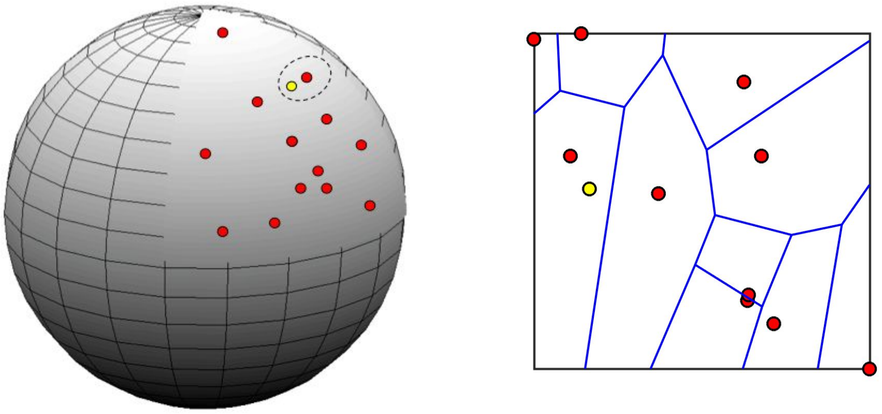

Figure 1. Many stored contexts and one current context. *The sphere shown here serves to represent a high-dimensional sphere where* m *stored contexts (components of matrix* W *) are illustrated by the red dots and one current state,* $\vec{r}$*, as a yellow dot (i.e. the vector representing the current state has the same magnitude and dimensionality as the stored contexts). The dashed line illustrates that one of the stored contexts is closest to the current state and hence this context would determine the output (e.g. motor output). The diagram on the right is a Voronoi tessellation of a 2D space illustrating the same idea. Each red dot is a 'seed' that generates a Voronoi region bounded by the blue lines.*

The current context, $\vec{r}$, will always be closest to one stored context $w_{(i,*)}$. The surface of the n -dimensional sphere of contexts can be considered as a Voronoi tessellation where each stored context is a 'seed' in the tessellation (Aurenhammer 1991). The 'recognised context', $w_{(k,*)}$, is associated with an output, $\vec{o}$, in the above examples that normally leads to a motor movement. If $\vec{o}$ leads to movement in the world, there is usually a new sensory input, a new motivational input and hence a new input vector, $\vec{r}$. Labels for $\vec{r}$ in each of these epochs could be $\vec{r}_{\mathrm{j}}$ and $_{\vec{r}\mathrm{j+1}}$, but we do not refer to them in the rest of the paper and simply discuss the fact that $\vec{r}$ changes, e.g. the 'path of $\vec{r}$ through sensory+motivational space', where $\vec{r} \in \mathbb{R}^{\mathrm{n}}$. Finally, there must be a mechanism to compare the new value of $\vec{r}$ with the expected value, sometimes called a 'motor error' (Waitzman et al. 1988; Porrill et al. 2004). We assume that in many cases there is a 'virtual' output, with no overt motor response, and a movement of $\vec{r}$ as a result but considering how that might be achieved is beyond the scope of this paper. Also, in the examples that follow (e.g. considering judgements of slant, depth, distance) we assume that there is only one Voronoi region determining the output (motor or otherwise) at any moment. We do not exclude the possibility that people may be engaged in more than one task simultaneously (e.g. walking and humming) and that this would be described better by more than one input vector and more than one Voronoi region.

We describe the input vector to the decision step, $\vec{r}$, as a 'sensory+motivational context'. By this we mean that $\vec{r}$ is a concatenation of a vector of sensory inputs (including different sensory domains), $\vec{s}$, and a vector of motivational inputs, $\vec{t}$, i.e.

$$\vec{r} = \vec{s} \,||\, \vec{t} \qquad \text{Eq (3)}$$

where $\vec{s} \in \mathbb{R}^{ns}$, $\vec{t} \in \mathbb{R}^{nt}$, $\vec{r} \in \mathbb{R}^{n}$ and, since $\vec{s}$ and $\vec{t}$ each add independent dimensions to $\vec{r}$, i.e. $n = ns + nt$.

4. A 2½D sketch

Marr and Nishihara (1978) put forward a hypothesis about how the brain might represent depth relationships in a scene despite the fact that the observer moves around. They introduced the term a '2½D sketch', a representation that was supposed to remain unaffected by vergence movements of the eyes. Neither Marr and Nishihara nor Marr in his book (Marr 1982) were very clear on the what the coordinates of this sketch might be and it has remained a tricky issue to define a useful coordinate system other than an image-based (retinal) coordinate frame or a 3D, world-based frame which are the principal coordinate frames used in computer vision (Hartley and Zisserman 2003; Davison 2003). Here, we describe an alternative version of a '2½D sketch' based on a graph of views connected by actions.

4.1 Looking around

If the eye (or a camera) were unable to translate in space and could only rotate about its optic centre (which is not quite true of the eye, but approximately so), then there would be a certain number of objects that it could fixate and a greater number of rotations (the square of the number of objects) that would take the eye from fixating one object to another. In most cases, these rotations will be saccades but large rotations of gaze may also involve head movements (and in the superior colliculus, these are often combined together so that the gaze change is the important parameter encoded, Corneil et al 2002). In a graph representation of this situation, each node would correspond to a context that includes (i) sensory data that helps identify the fixated object and (ii) a signal that specifies the next desired fixation point. Together, these are sufficient to specify the rotation of gaze (e.g. saccade) required to fixate the new target.

As we discussed above (Section 3), these two elements (sensory data and a signal about the next goal) could be concatenated into a long list or vector and compared to a large number of stored contexts that are in the same format (i.e. vectors in the same space). This type of representation is a graph rather than a reconstruction in a 3D coordinate-frame: the nodes are contexts and the edges joining the nodes are actions.

A practical issue concerns the number of stored contexts that would be required to make this work. Eye movement experiments suggest that humans store features in groups, which allows a hierarchical, coarse-to-fine strategy to be used to navigate between features and a consequent reduction in the number of relationships that would need to be stored (Findlay and Gilchrist, 1997; Watt 1987).

4.2 Moving around an object

The idea that object recognition in humans and animals is based on sets of 2-D views rather than 3-D internal models has a long history (Bülthoff and Edelman 1992; Tarr and Bülthoff 1998) and has been bolstered by recent comparisons between object recognition in infero-temporal cortex and in deep neural networks (DiCarlo et al 2012). Representing the 3D structure of an object rather than simply recognizing it in different poses involves more than a large store of 2D views. For a graph representation of object shape, just as in the case of gaze shifts in the previous section, the different views of the object form different nodes of the graph and these nodes are connected by edges: the actions that would turn one view into another. This time, assuming a static world and a moving observer, which is our assumption throughout this section on a 2½D sketch, the actions are head movements. Figure 2 shows some of the different views of a cube that can be obtained by a monocular observer moving in different directions while maintaining gaze on the object or, in the case of the highlighted pair of views, a binocular observer viewing the same cube. A sphere, on the other hand will always project a circular image. A planar slanted surface will project to an image that deforms in characteristic ways. In the latter case, the same images (nodes of the graph) connected by smaller head movements (edges) would be caused by, and would lead to the perception of, a surface that is more steeply slanted in depth (Glennerster 2016).

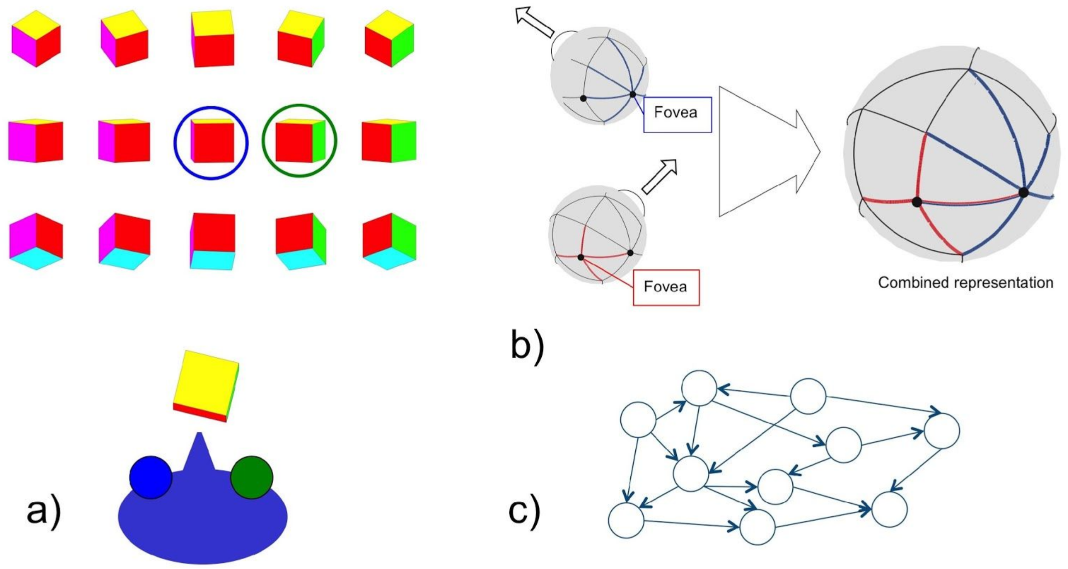

Figure 2. Views connected by actions. *a) The left and right eye's view of a cube are circled above and shown in plan view below. Also shown are many other views that a monocular observer would obtain if they moved their head while fixating the cube. b) Rotations of the eye (saccades) also give rise to a set of images that are related to one another in a consistent way (here, shown on a single sphere). c) For both a) and b) the images can be described as a graph of views connected by actions. b) is reproduced with permission from Glennerster et al, 2001. Figure is adapted from Glennerster (2016).*

Our proposal is that the observer's perception of a cube like the one shown in Figure 2 involves nothing more than movement of $\vec{r}$ between a set of Voronoi regions that are appropriate, so that for every movement the observer makes $\vec{r}$ lands up in the appropriate Voronoi region (i.e. the sensory information that observer receives matches the expectation for that movement). For example, if an observer received all the images shown in a) in a random order as they moved their head then they would not perceive a cube but if they received them in a way that was consistent with viewing a real cube as they moved then they would perceive a cube. However, the prediction of the sensory consequences does not have to be entirely accurate to perceive a static 3D object. For one thing, the degree of tolerance to error may depend on the task, as we discuss in Section 5.

4.3 Object distance

Finally, to include all the elements Marr and Nishihara (1978) envisaged in a 2½D sketch, the representation must incorporate information about the relative distance of objects. A monocular observer moving in different directions causes objects to move relative to one another in the image (motion parallax). While this can, in theory, be used to calculate the 3D layout of objects in the scene, an alternative is that the brain stores a graph of sensory+motivational states, connected by actions, just as in the case of 3D shape described above. In this case, the aspects of the images that are important are the angular separations of objects and how these change as the observer moves their head. Very distant points like the stars do not change in the representation as the observer moves. For example, if the current context is that the observer is looking at the pole star and they want to look at Ursa Major then Cassiopeia and then back to the pole star, the magnitude and direction of those three saccades, hence those nodes in the stored graph, will be independent of the location of the observer in the northern hemisphere. Provide that the angles of these saccades are large then the fact that the angles do not change with observer translation means that the objects must be distant. This type of information can be used to identify near and far objects in a vista space (Glennerster, Hansard and Fitzgibbon 2001; Glennerster 2016). It also means that distant objects 'anchor' the representation relative to the world.

4.4 Neither egocentric nor allocentric

There are many tasks that only make sense with respect to certain coordinate frames ('point to your left', 'look behind the cupboard', etc). The representation we have described here is capable of supporting actions defined by these relationships without having at any stage a 3D coordinate frame that describes the location of all points in the scene in an ego-, object- or world-centred frame. The representation is a graph of sensory+motivational states that carries information about the consequences of orienting movements (e.g. saccades) required to fixate different objects in the vista space and the translations of the head that would result in different views of a surface, where each sensory+motivational state is a node in the graph and the movements that would take the system from one state to another are the edges. Nevertheless, it is worth enunciating the links between this type of representation and one using an ego- or allocentric frame. If the observer is only making saccades (pure rotations of the eye), then the images and actions depicted by the nodes and edges in Fig 2c can be mapped onto the sphere shown on the right in Fig 2b. The eye is at the centre of this sphere and, in this sense, the representation for saccades is egocentric. But when we consider how distant points are represented, there is a sense in which the representation is also world-based, as discussed above for the example of looking between stars. All static points in an allocentric representation maintain their coordinates when the observer moves. In the representation we propose, the relationship between stars remains constant when the observer moves (either rotates or translates) and in this sense distant points provide an allocentric foundation for the representation. Similar representations based on points at infinity have been described in computer vision (Kumar *et al*, 1995). It may sound contradictory to say that a graph of sensory states has similarities with both ego- and allocentric representations at the same time but, since the representation has no 3D origin or coordinate frame, it is not identical to either of these types of 3D frame (Glennerster et al, 2001, 2009; Glennerster, 2013; Glennerster 2016).

5. Falsifiable predictions

It is always helpful to make clear what experimental results would falsify a theory. The focus in this paper is on putative 3D coordinate transformations that transform information from a visual reference frame into egocentric or allocentric frames. Convincing evidence that this chain of events occurs in the brain would make our proposal untenable. Here, we outline examples of neurophysiological and psychophysical results that would be highly problematic for our proposal (for more detail, see Glennerster, 2016). We also discuss task-dependent performance since, in the past, a finding of performance across a number of tasks that was consistent with a single (distorted) representation of space has been used to argue that the brain forms a multi-purpose 3D reconstruction of the scene. We question this conclusion in the light of more recent data.

An example of a psychophysical experiment that would challenge our hypothesis would be one showing evidence of a head-centred representation in situations where the task had nothing to do with head orientation. According to the reconstruction hypothesis, egocentric representations such as head- and body-centred representations are generated as an intermediate step to (or, at the very least, in addition to) building world-centred representations (Burgess and O'Keefe 1996; Burgess, 2008; Whitlock, Sutherland, Witter, Moser and Moser 2008). By contrast, if the brain simply uses the relevant sensory data for a particular task then there should be no psychophysical evidence of representation in an irrelevant coordinate frame. For example, if an observer fixates a point and there is a moving stimulus on peripheral retina then, after a while, there will be adaptation to the moving stimulus at that retinal location. Suppose the observer now rotates their head while maintaining fixation so that the moving stimulus now falls at a different head-centric location. Note that everything in retinal coordinates remains the same in this situation and also in spatiotopic (world-based) coordinates (Melcher, 2005; Knapen, Rolfs and Cavanagh 2009; Zimmerman et al 2011; Turi and Burr 2012) because the position of the eye in relation to the scene is unchanged. If we arrange it so that the adapting phase is before and the test is after the head rotation, then we can look for any trace of adaptation that applies in a head-centred frame. A hypothesis that suggested allocentric coordinates are inherited from head-centred coordinates (e.g. Snyder et al, 1998; Melcher, 2005) would predict adaptation in head-centred coordinates. By contrast, evidence of head-centred adaptation in this task would present severe difficulties for our hypothesis.

Mechanisms that have been proposed for implementing coordinate transformations in the brain depend on anatomical relationships such that neighbouring neurons in one area map to neighbouring neurons in the next. As a consequence, there is a consistent *format* to the new representation (e.g. Pouget et al 2002). Neurophysiological evidence of representations in particular formats could also present difficulties for our proposal. For example, it has been pointed out that if one signal, such as a head direction signal, is to determine the output of a transformation (e.g. transforming an egocentric representation into a world-centred one) then this could be done by generating multiple copies of the egocentric representation and using the head-direction information as a gating signal to determine which egocentric map passed on information to the world-centred map (Byrne et al, 2007). Although one can see how this might work for the case of rotation, it would be much harder to do the same thing for translation where there are many more potential input-output relationships. In any case, discovery of an anatomical mechanism of this sort for carrying out transformations, including duplication of input representations, would falsify our proposal since it is based on quite different principles. In Section 6.3, we discuss strong and weaker interpretations of the idea of an egocentric reference frame. An advantage of our proposal is that, at the level of neural mechanisms, we are not proposing any novel or complicated mechanism of the type put forward by Byrne et al (2007). The only neural operation we mention explicitly in our proposal is a fundamental and widely accepted one in neuroscience, namely the process of comparing a long vector of firing rates with a large number of similar length vectors stored as synaptic weights (Section 3).

Task-dependent performance is a crucial component of the argument about 3D reconstruction versus a graph of sensory+motivational states. Indeed, Gogel (1993) argued the opposite case to the one we are putting forward, pointing to the fact that performance in a number of different tasks could be explained by assuming by a similar distortion of visual space and using this 'explanatory parsimony' as evidence that the brain builds a single, multi-purpose internal representation of the scene. However, there is now considerable evidence that the internal representation used by the visual system is something much looser than this, with no single representation being capable of explaining performance. Instead, different tasks lead to the adoption of strategies or heuristics by the visual system (Koenderink et al, 2002; Smeets, Sousa and Brenner, 2009; Svarverud, Gilson and Glennerster 2012; Glennerster et al 1996; Knill, Bondada and Chhabra, 2011). Task dependency fits with the idea of recognising a sensory+motivational context: the task forms part of the motivational context and this dictates which aspects of the available sensory information are relevant. Together, these determine what the output should be. Although we are discussing task-dependency in a section on 'falsifiable predictions', we believe that this has already been tested and that, in the years since Gogel made his claim, a large body of evidence has built up against the idea of a single, multi-purpose representation of space.

6. Challenges

Next we consider some challenges to our proposal, starting with challenges arising from the philosophical literature.

6.1 Representation and action-based theories

Some philosophical work proposes alternatives to representation in 3D coordinate frames by emphasizing, as we do, the connection between spatial perception and action (O'Regan and Noë, 2001; Noë, 2004). We begin by distinguishing between our proposal and these related ideas, and explaining why a key challenge to those ideas does not apply to our proposal.

O'Regan and Noë introduce their theory as follows:

> We propose that seeing is a way of acting. It is a particular way of exploring the environment. Activity in internal representations does not generate the experience of seeing. The outside world serves as its own, external, representation. The experience of seeing occurs when the organism masters what we call the governing laws of sensorimotor contingency.
>
> O'Regan and Noë, 2001: 939

O'Regan and Noë propose that visual experience consists in a way of acting which involves the exercise of sensorimotor knowledge—knowledge of how sensory experience will change given certain movements—rather than in the reconstruction of a detailed representation of the visible scene. For example, visual experience of the shape of a cube consists in an active movement which exercises knowledge of the sensory consequences of this kind of movement around a cube. Similarly, according to our proposal, visual perception of the shape of a cube consists in movement and a matching of current and stored sensory+motivational contexts which registers the sensory consequences of this kind of movement around a cube. In both theories, action across a space replaces the need for an internal 3D model of that space.

However, there is at least one important respect in which O'Regan and Noë's theory goes further than what we propose here: they argue that the process of vision consists in a way of acting, rather than a process involving internal visual representations. By contrast, though we deny that the process of vision consists in transformations between 3D coordinate frame representations, we propose a process that fundamentally involves internal visual representations of a different kind. More specifically, we propose that a large amount of information is stored and that this store of information is exploited in a process of comparing current and stored sensory+motivational contexts. This store of information is *W* in the formalism above. Not every brain state which carries information is a representation of that information: there is a good sense in which any effect carries information about its causes, and it is notoriously difficult to say exactly what it takes for an information-carrying state to count as a representation. But by most criteria *W* is a very plausible candidate for representation, because of the simple computational role it plays in perception and action.

For example, one common criterion for a brain state's representing, rather than merely carrying, information is that carrying this information is the brain state's *function*. That is, roughly, either the ontogenetic (Dretske, 1981) or the phylogenetic (Millikan, 1984) explanation of why the brain state occurs in the organism lies in a functional role which the brain state plays in virtue of carrying the information in question. The brain state corresponding to *W* carries information about various sensory+motivational contexts which might obtain, and in virtue of carrying this information, this brain state plays a distinctive computational role: elements of *W* are matched to the actual sensory+motivational contexts marking any given episode of perception. Our hypothesis is that this computational role is integral to successful perception and action. Successful perception and action are, in turn, integral to the organism's development and survival. So it is very plausible that *W*'s functional role as a store of information about sensory+motivational contexts explains why the brain state corresponding to *W* occurs in the organism. It is very plausible that by this common criterion for representation, *W* is a representation. (In the next section, we assess in more detail what both $\vec{r}$ and *W* represent.)

Because *W* is a very plausible candidate for representation, one influential criticism of O'Regan and Noë's theory does not apply to our proposal. A theory is behaviorist--in the standard philosophical terminology--if it reduces cognition to behavior, or if the fully articulated theory eliminates references to cognition, replacing them with references to action or behavior. Behaviorism is notoriously problematic, in part because it is too liberal in counting as genuinely cognitive any system which exhibits the relevant behavior in the relevant circumstances. Ned Block (2001) criticizes O'Regan and Noë's sensorimotor theory on the grounds that it commits to this problematic aspect of behaviorism: it counts as having visual perception any system which acts in the right way in the right circumstances, whatever (if anything) in the system mediates between circumstances and action. It is a delicate issue whether O'Regan and Noë's sensorimotor theory is in fact behaviorist in this way. But we need not resolve that issue in order to see that our proposal is not behaviorist and does not fall foul of Block's criticism: our proposal treats visual perception as fundamentally based on certain stored representations, and so does not claim that any form of behavior is by itself sufficient for visual perception.

Our proposal does deny that the process of vision consists in computation over representations *of a certain kind*: spatial representations specified in terms of 3D coordinate frames. So challenges to our proposal lie not in the idea that internal representations are required to explain perception, but in the more specific idea that 3D coordinate frame representation, or at least some form of representation inconsistent with our positive proposal, is required to explain perception. We turn to this next.

6.2 Perceptual systems vs mere sensorimotor systems

The challenges we assess in this section argue that our proposal is ill-placed to explain genuine perception of the environment. We have described a system through which the environment's sensory effects generate motor responses, but the challenges we assess here maintain that we have not thereby explained genuine perception of the environment. There are difficult philosophical questions about what, in principle, constitutes genuine perception, and of course difficult empirical questions about how genuine perception is achieved by the human brain. We do not pretend to offer a full explanation. Instead we argue that our proposal is as well-placed in this respect as a 3D coordinate frame theory.

Start with the idea that perceptual representation is distinct from mere sensation, in that perceptual representation is representation of a distal environment (Burge, 2010). One possible challenge to our proposal argues that it cannot do justice to this distinction.

The sensory component of a current sensory+motivational context is often identified above as an 'image'. It may be tempting to read this as if it were an appeal to sensory impressions or sense data, construed as proximal objects of conscious experience (Russell 1912; Price 1950; Moore 1953). If this is combined with a similar reading of the motivational component of a sensory+motivational context, as a felt urge or motivating sensation, our proposal may seem to commit to a *phenomenalist* scheme of representation--a scheme such that the world represented by visual processes consists in nothing more than a series of experienced sensations connected by transitions which those sensations motivate. Our proposal should not be read in that way. The sensory component of a current sensory+motivational context includes images only in the sense that the set of neural firings $\vec{r}$ may include neural firings corresponding to the retinotopic signals that we know about in visual areas V1, V2, V4, V5, MST, IT etc. These signals form part of the input to a process that is responsible for 3D perception, just as in a coordinate frame theory, retinotopic signals form part of the input to the computations required to construct 3D coordinate frames. In our proposal, just as in a coordinate frame theory, retinotopic signals need not be construed as objects of conscious experience. Our proposal does not entail a phenomenalist scheme of representation.

In fact our proposal is as well-placed as a coordinate frame theory to explain perception of distal vista space. Take a coordinate frame theory in which the format of representations in the visual cortex resembles the spatial layout of distal vista space: spatial relations among units of representation in the cortex correspond to spatial relations among the objects represented (e.g.

Byrne et al., 2007). This resemblance does not explain representation of the distal spatial layout. Unlike representation, resemblance is a symmetrical relation, and the environment does not represent the cortex. Moreover, many patterns which resemble the layout of an environment do not represent it. For example if the layout of objects on my desk happens to resemble the layout of objects in your vista space, neither represents the other, because there is no appropriate causal or explanatory connection between them (Putnam, 1981). It is controversial what *does* suffice for a brain state to represent a feature of the environment (Stich and Warfield, 1994) but on the face of it, any plausible explanation of this achievement will have to appeal to causal or explanatory relations between features of the environment, brain states which carry information about those features, and/or brain states' functional contributions to action on those features.[3] For example, a coordinate frame theory might claim that the spatial layout of vista space systematically explains the corresponding spatial arrangement of neuron-firing in the cortex, and that this relationship in turn systematically explains the observer's capacity to act on vista space; that is why the relevant brain states constitute representation of your vista space, while the layout of objects on my desk does not.

The same resources are available if our proposal is correct. In the first instance, the set of synaptic weights W carries information about sets of possible neural firings. But W and $\vec{r}$ also carry information about distal features of the environment including its spatial layout, a layout which systematically explains the various values of $\vec{r}$ as the observer moves their head or eyes, and so systematically explains why $\vec{r}$ is matched with certain elements of W during that episode. In turn, these matches systematically explain the actions on vista space which are outputs of the process. For example, take a scene in which three objects are relatively close to the observer. The distance

[3] Some but not all philosophical theories of what a brain state represents appeal to the brain state's functional contributions to action. As discussed in Section 6.1, some theories appeal to the idea that a representation of *X* has the function of carrying information about *X*. The representation's having this function will involve its functional contribution to adaptive actions involving *X*. By contrast, Fodor (1987) argues that what a brain state represents turn only on what causes or would have caused that brain state, not on the brain state's functional contribution to action.

to these objects explains why, as the observer moves her head in different directions, the angular separation between the projection of the objects' on the retina changes (see *Object distance* above). This contribution to $\vec{r}$, over the course of the visual episode, marks a distinctive series of sensory+motivational contexts precisely in that it is a consequence of the objects' being relatively close (if all three points were distant there would be little change). So when $\vec{r}$ is matched with elements of *W*, yielding further actions distinctive of the sensory+motivational contexts, those distinctive further actions occur *because* the objects are close--more specifically, they occur because $\vec{r}$ and elements of *W*, and the process of matching them, carry information about the objects' being close.

This and similar examples (Glennerster, Hansard and Fitzgibbon 2001, 2009; Glennerster 2013) give us reason to think that the mechanism we have described represents the spatial layout of the observer's distal environment. The reason we have given is in line with Tyler Burge's (2010) criterion for perception of a distal environment. For Burge, perceptual constancy -- i.e. the use of varying sensory consequences to generate information about constant features of the distal environment -- is the key criterion for perception of a distal environment. We have explained how, on our proposal, information about distal layout is extracted from variations in $\vec{r}$.

We do not insist that Burge's criterion is in fact sufficient for genuine perception. We simply note that our proposal is as well placed in this respect as a theory according to which constant features of the distal environment are represented in a 3D coordinate frame. We also leave it open which specific elements of the mechanism we have proposed bear representational content about the distal layout of vista space. On an 'embodied' approach to our proposal, the content-bearing event is the entire loop, including the physical movements which take the observer between different values of $\vec{r}$ as well as the internal process of matching $\vec{r}$ with elements of *W*. Alternatively, on an approach which treats only internal brain events and brain states as bearers of representational content, the content-bearing event or state might be any or all of $\vec{r}$, elements of W, and the internal process of matching them; it might also include a 'motor error' signal indicating divergence between expected and actual values of $\vec{r}$ (Section 3; cf. Clark 2013).

Now consider a further challenge, which is posed by a more demanding characterization of perception. According to this more demanding characterization, the form of perception enjoyed by humans is *objective* in the following sense: we perceive features of the distal environment which are in principle independent of the observer's actual and possible responses.[4] The challenge maintains that our proposal cannot explain how perception of such features occurs, and so cannot explain the form of perception enjoyed by humans. As before, our response is not to offer a full explanation of how this occurs, but to argue that our proposal is as well-placed in this respect as a 3D coordinate frame theory.

John Campbell illustrates this more demanding characterization of perception as follows:

> Gibson seems to have thought that the affordances provided by an object are all that we ever see. This view is hard to sustain. Once, when I visited Warwick University Psychology Department, someone told me the following story. One year, pigeons started nesting in the concrete interstices of the multi-storey car parks in the university. Presumably, according to my informant, pigeons perceive these Gibsonian affordances directly. They would immediately look like good nesting places. But it had never occurred to humans that it was so. These affordances are not perceived by humans. But it is not as if the matter is entirely opaque to us. For though we cannot perceive the affordances directly, we can see the reasons why the interstices would be good for nesting. We do not see the affordance itself; we see the ground of the affordance.

[4] The phrase 'objective perception' might understood in various ways. Here we stipulate that it is to be understood as we just defined it, and as illustrated in the next paragraph.

Campbell, 2002: 143

The ground of an affordance is an intrinsic feature of the environment which explains that affordance. For example, the shapes of the concrete interstices explain why the interstices make possible certain actions, such as nesting, and certain experiences, such as those enjoyed by someone looking at the interstices. The grounds explain why these forms of action and perception are possibilities, but are in principle independent of these possibilities: if agents and observers were differently constituted, the same shapes would not afford the actions and experiences which they actually afford. So perceiving the grounds of affordances is objective perception in the sense defined above.

Campbell claims that humans perceive the grounds of affordances rather than, or in addition to, perceiving the affordances. Now if we perceive affordances as Gibson thought (see also Nanay, 2015), our proposal is well-placed to explain how we do so. This is because, according to our proposal, the spatial layout of vista space is visible insofar as iterated matches of $\vec{r}$ with *W*, and the movements taking an observer between them, carry information about the spatial layout of the environment; in the first instance, this information about the environment is information about possibilities for active manipulation of visual stimulation which the environment affords. But for the same reason, a natural challenge to our proposal argues that it is ill-placed to explain perception of the grounds of these affordances, as opposed to the affordances themselves.

One approach to this challenge insists that, in fact, the perceptible grounds of affordances are just more affordances. For example, the shapes of Campbell's concrete interstices are perceptible by humans in terms of the human actions which move an observer between various values of $\vec{r}$. More generally, this approach insists, all perception of spatial properties represents those properties in terms of the observer's actual and possible movements (Poincare, 1946). However, to address the challenge at its most demanding, suppose that there is objective perception of space, in the sense that we perceive spatial features which are in principle independent of their consequences for perception and action. Any explanation of the capacity for objective perception that is consistent with our proposal will have to start from the fact that information about the environment's affordances for action and perception is also information about the grounds of those affordances. This may seem unsatisfactory as a basis for explaining objective perception, but in fact a coordinate frame theory is in just the same position.

For example, consider how objective perception might be explained by a coordinate frame theory which appeals to representations with a spatial format in the cortex, corresponding to the spatial layout of vista space. In a theory of this kind, the spatial arrangement of neurons--one ground of the representation's functional contribution to spatial perception and action--resembles the spatial layout of the environment--the grounds of the environment's affordances for spatial perception and action. But as we saw above, this resemblance does not by itself explain representation of the environment's spatial layout; a brain state represents a feature of the environment only insofar as there are appropriate causal or explanatory connections between the feature represented, the brain state that represents it, and/or the brain state's contributions to action. That is, even on a theory which postulates cortical representations with a spatial format, the environment's spatial layout is represented only insofar as that layout provides affordances for action and perception. For example, the shapes of Campbell's concrete interstices are represented only insofar as these shapes provide affordances for action on them and perception of them. More generally, the grounds of affordances are represented only insofar as information about the affordances is also information about their grounds. A theory which postulates cortical representations with a spatial format is no better placed than our proposal, when it comes to explaining objective perception.

Not every coordinate frame theory will appeal to a spatial *format* in cortical representations. Some coordinate frame theories may propose other reasons for saying that brain states have *contents* expressible in a 3D coordinate frame (Grush 2000). But as we have seen, the reasons for saying that brain states have such contents must lie in appropriate causal or explanatory connections between the spatial relations which constitute the 3D coordinate frame, the brain states that represent those relations, and/or the brain states' contributions to action. So again, a brain state's coordinate frame contents would include the grounds of affordances for action and perception only insofar as those grounds provide affordances for action and perception. A theory which postulates representations with a coordinate frame content is no better placed than our proposal, when it comes to explaining objective perception.[5]

It is often suggested that the key to objective perception is representation in an allocentric, rather than egocentric, coordinate frame (O'Keefe and Nadel; Evans, 1982; Grush 2000).[6] We now briefly explain why representation in an allocentric coordinate frame is no better placed than our proposal, when it comes to explaining objective perception.

An allocentric, as opposed to egocentric, coordinate frame is a coordinate frame whose origin and axes are independent of the observer's current position. For example, O'Keefe's slope/centroid model (1991) is constructed as follows. By keeping track of its own movements, an organism learns the egocentric vectors from its current position and orientation to various landmarks. This allows the organism to compute, in addition, egocentric vectors between landmarks. The *slope* of the environment is the average of the gradients of these vectors. The *centroid* is the 'centre of mass' of landmarks in the environment, for example the average of the egocentric vectors to

[5] For similar reasons, Campbell (2002) claims that objective perception must be a non-representational form of experience. Our position here is neutral about this claim.

[6] Things here are complicated by the fact that O'Keefe and Nadel take a cognitive map in the hippocampus to be the locus of allocentric and objective representation, while perception-specific systems operate in egocentric reference frames. However, the points which follow are independent of this particular view.

landmarks. The slope and the centroid are invariant with translations of the organism's current position, so allocentric vectors to goals can be defined in terms of distance from the centroid and angle from the slope.

As we defined it, objective perception is perception of features which are in principle independent of the observer's actual and possible experiences and actions. The claim that perception represents spatial layout in terms of allocentric vectors, such as those defined by O'Keefe's model, does not explain how objective perception is possible. In the first instance, vectors are introduced into O'Keefe's model insofar as the organism keeps track of its current movements. The model then explains how the organism's representations abstract from its current position and movements. But that is not the same thing as abstracting from all possible positions and movements of the organism. For example, if representation of a vector was introduced into the model as representation of the organism's current movement between landmarks, the abstraction generates, at best, representation of *some* movement of the organism between landmarks (movement which is not currently available to the organism). At least, nothing in the model explains how a further abstraction is possible, such that the spatial features represented are independent of the organism's possible experiences and actions (Campbell 1993).

Consistent with this last point, Grush (2000) gives empirical reasons for thinking that O'Keefe's model should be understood in terms of a capacity for offline or imaginative adoption of an 'alter-ego-centric reference frame': the organism represents allocentric vectors by imagining occupying a point of view other than its current one. Grush holds that, in the first instance, perception represents spatial layout insofar as perceptual representations contribute to skilled action; perceptual representations abstract from current affordances for action insofar as perceptual representations are integrated with the offline or imaginative capacity. On this approach, it is especially clear that the abstraction generates, at best, representation of some possible affordances (those which would be available from a point of view other than the organism's current one). Grush suggests that this amounts to a form of objectivity--a way of representing one's current location explicitly, as one location among many, as opposed to representing one's current location only implicitly, as the origin of an egocentric frame. However, it is not clear why this form of objectivity should require an allocentric coordinate frame, rather than a kind of offline or imaginative occupation of a point of view that is consistent with our proposal. (In the terms of our proposal, this might involve offline or imaginative changes in $\vec{r}$, but discussion of imagination, rather than perception, is beyond the scope of this article.) Moreover, Grush's limited form of objectivity does not suffice for objective perception as we defined it here. The deeper problem of how objective perception is achieved is not solved by postulating allocentric coordinate frames, any more than it is solved by our proposal.

In summary, a coordinate frame theory is no better off than our proposal when it comes to explaining either genuine perception, as opposed to mere sensorimotor responses, or objective perception. This can be obscured by the fact that our proposal appeals immediately to action in a space, as the observer's means of perceiving a 3D layout. But in fact, even on a coordinate frame theory, a brain state can be said to represent features of the environment only insofar as there are appropriate causal or explanatory relations between these features, the brain state, and the brain state's contributions to action. The same resources are available on our proposal.

6.3 Challenges in relation to neuroscience

6.3.1 Where is W?

We have described the set of stored contexts, *W*, in a very general way since the location of these synaptic weights is not critical to the argument. To illustrate this point we describe here two examples of a motor response to a sensory+motivational context. In one, the critical neurons are in the spinal cord while in the other the fluctuation of firing rate of a single neuron in the cortex moves $\vec{r}$ from one Voronoi region to another. In a patellar reflex, the most important afferents that determine the sensory context are from stretch receptors in the quadriceps muscle and the critical synapses are with alpha motor neurons in the spinal cord, although the briskness of the reflex can be modulated by input from higher levels in the nervous system. These stretch receptor afferents can be considered to be part of $\vec{r}$ and the effectiveness of the spinal synapses is recorded in W. Described in terms of our proposal, a patellar reflex corresponds to an extremely large Voronoi region because the same output is triggered by a very large range of values of $\vec{r}$. (Since $\vec{r}$ is a vector and especially as it always has the same length, it might be more appropriate to talk of the 'direction' rather than 'value' of $\vec{r}$, but for the sake of simplicity, we have used the term 'value' throughout.) In a quite different example, the firing rate of a single neuron in V5/MT could move $\vec{r}$ from one Voronoi region to another, reversing the perceived direction of motion of a rotating cylinder (Dodd et al, 2001). On the critical trials, the direction of motion of the cylinder is ambiguous. There is debate about how many neurons might need to vary their firing rate to cause (or reflect, Nienborg and Cumming, 2009; Cumming and Nienborg, 2016) a change in perception but it is clear that fluctuations in the firing rate of a single neuron are highly predictive of the animal's choice. This means that a tiny change in the value of $\vec{r}$ causes it to cross the boundary between two Voronoi regions. The elements of $\vec{r}$ that are common to both regions define the context (the animal is doing an experiment, it knows it has to maintain fixation and attend to the direction of motion of the cylinder, etc) while the firing of only a tiny number of neurons (in theory, even a single neuron) determine whether the animal perceives the cylinder as rotating left or right. In this example, a large range of different inputs contribute to the decision (mostly defining the context but a few determining the perceived rotation of the cylinder as clockwise or anticlockwise) and these need to be brought together in one place. Where this might happen in the brain is up for discussion. One possibility might be the cerebellum which receives input from all over the cortex, has a much more significant input than the cortico-spinal tract, is able to compare long input vectors with large numbers of vectors of stored weights (Marr, 1969; Albus, 1971; Eccles et al, 1967; Manzoni, 2005) and is known to use multi-modal contexts to control motor output. There is also a considerable literature on the possible role of the cerebellum in cognition (Schmahmann, 2004). Other regions that have a wide range of both sensory and motivational inputs and that may be relevant to consider in any discussion of a biological implementation include the basal ganglia (Krauzlis et al, 2014) and the prefrontal cortex (Mante et al, 2013).

6.3.2 Coordinate transformations (or not) in posterior parietal cortex

In many accounts of coordinate transformations, posterior parietal cortex plays a special role as the site of different ego-centric coordinate frames and as a way-station in the process of transforming visual information into an allocentric frame. Here, we take the example of data that has been cited as evidence of a 'hand-centred' or 'head-centred' representation of visual space and consider it in relation to two extreme interpretations of this notion. For example, 'hand-centred' neurons respond selectively to the direction and distance of a target object from the hand (Buneo and Andersen, 2006) and 'gaze-direction' neurons respond best when a target is at a given retinal location and the eye has a certain position with respect to the head (Zipser and Andersen, 1988; Pouget et al, 2002). Taken literally, transforming 3D visual information to a hand-centred frame could mean the following. Suppose that an observer holds up their hand so that it is in front of their face. Then, two axes of the hand-centred frame would map onto two axes of the visual reference frame (x_{vis} and y_{vis} map onto x_{hand} and y_{hand}, say) while the third visual axis, derived from the disparity tuning of neurons in V1, say, maps onto the third axis of the hand-centred frame (z_{vis} maps onto z_{hand}). Now consider what happens to the representation of the entire visual scene when the hand rotates to be palm up. z_{vis} now maps onto y_{hand} while y_{vis} maps onto z_{hand}. An intermediate hand position would require some complex mix of disparity sensitivity and local sign information to determine which V1 firing rates were transferred to which neurons in the hand-centred frame. The process would be even more complicated if the hand translated to a different location because then the depth of scene points would need to play a role in calculation. But in any case, all this must happen rapidly as the hand moves and, by assumption, it applies to the whole visual field. The processes would generate a representation of the whole visual scene and would follow the format of a visual representation such as that in V1 but now centred on the hand and oriented according to the pose of the hand. Data supporting such a literal transformation in the brain would be strong evidence against our proposal.

A much less extreme version of an 'ego-centred representation' is when a neuron responds to a combination of sensory cues that are relevant to a potential movement. In ventral premotor cortex, for example, which is involved with the preparation of movement, some neurons have both a tactile receptive field on the arm and a visual receptive field but the neuron responds most strongly when the visual stimulus comes towards the tactile receptive field. It does this despite changes in a variety of other factors such as the animal's eye position, arm position relative to the body or of arm position relative to the scene (Graziano et al, 1994). Responses with these characteristics reflect sensitivity to a particular combination of stimuli (including the relative visual direction between the hand and the target) but it is quite possible that that is *all* they need to do, rather than being a stepping stone in a coordinate transformation as described above. Of course, neurons that are sensitive to a combination of inputs from two modalities are common *(*Zipser and Andersen, 1988; Pouget et al, 2002). Pouget et al have shown how such neurons could be involved in a coordinate transformation (Fig 3) but their existence does not, by itself, support the notion of a true 3D coordinate transformation. Indeed, this type of logical 'AND' operation is useful for reducing redundancy in the output from the cortex and is widely recognized as an important cortical operation without being part of coordinate transformations (Barlow, 2001; Földiak, 1990).

6.3.3 Processing of sensory input to generate $\vec{r}$

In describing our proposal, we have said little about the role of the cerebral cortex. In particular, we have claimed that at the stage $\vec{r}$ is compared to the rows of *W*, $\vec{r}$ does not consist of 3D coordinates in an ego- or allocentric frame derived from visual, retinotopic coordinates by a 3D transformation process. Note that the proposal that $\vec{r}$ is compared to *W* does not, by itself, rule this out. Almost any theory of the brain can be couched in terms of $\vec{r}$ and *W*, including one in which $\vec{r}$ lists the 3D location of the observer and various objects in the scene while each row of *W* consists of a similar list of 3D coordinates. We have suggested that the cortex has an important role in redundancy reduction, but this is a long way from the type of processing that would be required to transform the representation of objects between different 3D coordinate frames.

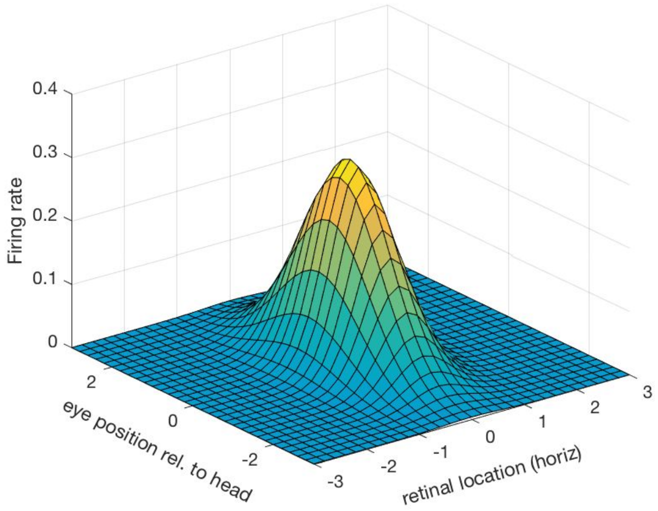

Figure 3. Neural response to a combination of cues (adapted from Pouget et al 2002). *The firing rate of a putative neuron is shown. The neuron has a retinal receptive field but the gain of the response is modulated by eye position relative to the head. As a result it responds best when*

the stimulus is in a certain retinal location AND *the eye has a particular position relative to the head. Units are arbitrary.*

6.3.4 Learning W

We have assumed in our discussion that *W* is already learned. A proper consideration of how that might occur (and how large *W* might need to be) is beyond the scope of this paper but it is clear that the learning of *W* presents a serious challenge. Recent developments in computer vision are relevant, particularly end-to-end learning of tasks in a 3D environment such as grasping objects (Pinto and Gupta, 2016) or navigating from one place/image to another (Zhu et al, 2016). These algorithms do not represent the environment using a 3D coordinate frame and Zhu *et al* learn using a combination of signals about the current image and the goal image (analogous to $\vec{s}$ and $\vec{t}$ contributing to $\vec{r}$ in our proposal). Classification in deep artificial neural networks has been described as an input feature vector being compared with many stored feature vectors in 4096-dimensional continuous space, where the similarity is measured as correlation between the current and the stored feature vectors (Krizhevsky et al, 2012). This is similar to Equation (2) above and to the idea that $\vec{r}$ will always fall within a Voronoi region around a stored vector, *w(i,*)*.

7. Conclusions

We have proposed that perception and action are achieved through a process of comparing a vector, $\vec{r}$, with a weight matrix, *W* (a long look-up table), to choose an output and that in order for perception and action to be achieved in this way, representation in 3D coordinate frames is not required. One advantage of our proposal is that it appeals to neural operations that are well accepted (the comparison of a vector of firing rates with many similar vectors stored as synaptic weights), whereas there are few specific proposals and no convincing evidence for neural operations that carry out the type of 3D coordinate transformation that would turn a visual 3D representation into an egocentric and then an allocentric 3D representation. Almost all theories of the brain could be cast in terms of generating a vector $\vec{r}$ that is then compared to a set of stored contexts, *W*, but we claim that the computations leading to the generation of $\vec{r}$ are relatively simple and, in particular, that they do not include the calculation of 3D coordinate transformations. We have argued that when it comes to meeting the philosophical demands on a theory of perception, a 3D coordinate frame theory is no better placed than our proposal. The real test of our proposal is in its empirical predictions and we have set out some here. But in future, it may be computer vision that sets the agenda, providing new hypotheses about possible representations of a 3D world and how to interact with it.

References

Albus, J. S. (1971). A theory of cerebellar function. *Mathematical Biosciences* 10(1), 25-61.

Andersen, R. A., Snyder, L. H., Bradley, D. C., & Xing, J. (1997). Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience* 20(1), 303-330

Attwell, D., & Laughlin, S. B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism*, 21(10), 1133-1145

Aurenhammer, F. (1991). Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)* 23(3), 345-405.

Barlow, H. (2001). Redundancy reduction revisited. *Network: computation in neural systems*, 12(3), 241-253;

Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S., & Movshon, J. A. (1996). A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Visual Neuroscience* 13(1), 87-100.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences* 89(1), 60-64.

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

Burgess, N., & O'Keefe, J. (1996). Neuronal computations underlying the firing of place cells and their role in navigation. *Hippocampus* 6(6), 749-762

Byrne, P., Becker, S., & Burgess, N. (2007). Remembering the past and imagining the future: a neural model of spatial memory and imagery. *Psychological Review* 114(2), 340.

Campbell, J.J. (2002). *Reference and Consciousness*. Oxford: Clarendon Press.

Camp, E. (2007). Thinking with maps. *Philosophical Perspectives* 21 (1):145–182.

Campbell. J.J. (1993). The role of physical objects in spatial thinking. In N. Eilan, R. McCarthy, and B. Brewer (Ed.s), *Spatial Representation*. Oxford: Blackwell.

Campbell. J.J. (2002). *Reference and Consciousness*. Oxford: Clarendon Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36 (3):181-204.

Cohen, Yale E., and Richard A. Andersen. (2002). A common reference frame for movement plans in the posterior parietal cortex. *Nature Reviews Neuroscience* 3: 553-562.

Colby, C. L. (1998). Action-oriented spatial reference frames in cortex. *Neuron* 20(1): 15-24.

Corneil, B. D., Olivier, E., & Munoz, D. P. (2002). Neck muscle responses to stimulation of monkey superior colliculus. II. Gaze shift initiation and volitional head movements. *Journal of Neurophysiology* 88(4): 2000-2018.

Cumming, B. G., & Nienborg, H. (2016). Feedforward and feedback sources of choice probability in neural population responses. *Current opinion in Neurobiology* 37: 126-132.

Cummins M and Newman P (2008) FAB-MAP: probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research* 27(6): 647–665

Davison, A. J. (2003). Real-time simultaneous localisation and mapping with a single camera. In *Ninth IEEE International Conference Proceedings: Computer Vision*..

Dehaene, S. & Changeux, J. P. (2004) Neural mechanisms for access to consciousness. In *The Cognitive Neurosciences III*, ed. M. Gazzaniga, Cambridge, MA: MIT Press: 1145– 58.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition?. *Neuron* 73(3): 415-434.

Dodd, J. V., Krug, K., Cumming, B. G., & Parker, A. J. (2001). Perceptually bistable three-dimensional figures evoke high choice probabilities in cortical area MT. *The Journal of Neuroscience* 21(13): 4809-4821.

Dretske, F. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Eccles, J. C., Ito, M., & Szentágothai, J. (1967). The mossy fiber input into the cerebellar cortex and its inhibitory control by Golgi cells. In *The cerebellum as a neuronal machine* (pp. 116-155). Springer Berlin Heidelberg.

Eilan, N., McCarthy, R. and Brewer, B. (Ed.s) (1993). *Spatial Representation*. Oxford: Blackwell.

Evans, G. (1982). *The Varieties of Reference*. Oxford: OUP.

Findlay, J. M., & Gilchrist, I. D. (1997). Spatial scale and saccade programming. *Perception* 26(9): 1159-1167.

Földiak, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*, 64(2), 165-170.

Fodor, J. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT/Bradford.

Franz, M. O., Schölkopf, B., Mallot, H. A., & Bülthoff, H. H. (1998). Learning view graphs for robot navigation. In *Autonomous Agents*, Springer US: 111-125.

Gillner, S., & Mallot, H. A. (1998). Navigation and acquisition of spatial knowledge in a virtual maze. *Journal of Cognitive Neuroscience* 10(4): 445-463.

Glennerster, A., Hansard, M. E., & Fitzgibbon, A. W. (2001). Fixation could simplify, not complicate, the interpretation of retinal flow. *Vision Research* 41(6): 815-834.

Glennerster, A. (2016). A moving observer in a three-dimensional world. *Phil. Trans. R. Soc. B* 371(1697): 20150265.

Graziano, M. S., Yap, G. S., & Gross, C. G. (1994). Coding of visual space by premotor neurons. *Science*, 266(5187), 1054-1057

Gogel, W. C. (1993). The analysis of perceived space. Advances in psychology, 99, 113-182.

Grush, R. (2000). Self, world and space: the meaning and mechanisms of ego- and allocentric spatial representation. *Brain and Mind* 1: 59-92.

Hartley, R., & Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. Cambridge: Cambridge University Press.

Jay, M. F., & Sparks, D. L. (1987). Sensorimotor integration in the primate superior colliculus. I. Motor convergence. *Journal of Neurophysiology* 57(1): 22-34.

Knill, D. C., Bondada, A., & Chhabra, M. (2011). Flexible, task-dependent use of sensory feedback to control hand movements. *Journal of Neuroscience*, 31(4), 1219-1237.

Koenderink JJ, van Doorn AJ, Kappers AM, Lappin JS. (2002) Large-scale visual frontoparallels under full-cue conditions. *Perception*, 31, 1467 – 1476

Krauzlis, R. J., Bollimunta, A., Arcizet, F., & Wang, L. (2014). Attention as an effect not a cause. Trends in cognitive sciences, 18(9), 457-464.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*: 1097-1105.

Kumar, R., Anandan, P., Irani, M., Bergen, J., & Hanna, K. (1995). Representation of scenes from collections of images. In *Representation of Visual Scenes*, *Proceedings IEEE Workshop* (pp. 10-17)

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature, 503(7474), 78-84.

Manzoni, D. (2005). The cerebellum may implement the appropriate coupling of sensory inputs and motor responses: evidence from vestibular physiology. *The Cerebellum*, 4(3), 178.

Marr, (1969) A theory of cerebellar cortex. *Journal of Physiology* 202: 437–470.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: WH Freeman and Company.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115-133.

Meilinger, T. (2008). The network of reference frames theory: A synthesis of graphs and cognitive maps. In *International Conference on Spatial Cognition.* Berlin / Heidelberg: Springer: 344-360.

Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural networks,* 9(8), 1265-1279)

Milford M and Wyeth G (2008) Mapping a suburb with a single camera using a biologically inspired SLAM system. *IEEE Transactions on Robotics* 24(5): 1038–1053

Millikan, R. (1984). *Language, Thought and other Biological Categories*. Cambridge, Mass.: MIT Press.

Moore, G. E. (1953). *Some Main Problems of Philosophy*, London: George, Allen and Unwin.

Nanay, B. (2015). *Between Perception and Action*. Oxford: OUP.

Nienborg, H., & Cumming, B. G. (2009). Decision-related activity in sensory neurons reflects more than a neuron's causal effect. *Nature* 459(7243): 89-92.

Noë, A. (2004). *Action in Perception*. Cambridge, MA: MIT Press.

O'Keefe, J. (1991). The hippocampal cognitive map and navigational strategies. In J. Paillard (Ed.), *Brain and Space.* Oxford: OUP.

O'Keefe, J. and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Oxford: OUP.

O'Regan, J. K. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences* 24: 939-1031.

Peacocke, C. (1992). *A Study of Concepts*. Cambridge, MA; MIT Press.

Pinto, L., & Gupta, A. (2016). Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *Robotics and Automation (ICRA)*, IEEE, pp. 3406-3413)

Poincare, H. (1902). *La Science et l'Hypothèse*. Paris: Flammarion.

Pouget, A., Deneve, S., & Duhamel, J. R. (2002). A computational perspective on the neural basis of multisensory spatial representations. *Nature Reviews Neuroscience* 3(9): 741-747.

Porrill, J., Dean, P., & Stone, J. V. (2004). Recurrent cerebellar architecture solves the motor-error problem. *Proceedings of the Royal Society of London-B* 271(1541): 789-796.

Price, H. H., (1950). *Perception* (2nd edition). London: Methuen.

Prince, S. J. D., Pointon, A. D., Cumming, B. G., & Parker, A. J. (2002). Quantitative analysis of the responses of V1 neurons to horizontal disparity in dynamic random-dot stereograms. *Journal of Neurophysiology* 87(1): 191-208.

Putnam, H. *Reason, Truth and History*. (1981). Cambridge: Cambridge University Press.

Rolls, E.T., and Treves, A. (1998). *Neural networks and brain function* (Vol. 572). Oxford: Oxford University Press.

Russell, B. (1912). *The Problems of Philosophy*. New York: Henry Holt & Company.

Saxe, R. and Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". *Neuroimage* 19: 1835-1842.

Schmahmann, J. D. (2004). Disorders of the cerebellum: ataxia, dysmetria of thought, and the cerebellar cognitive affective syndrome. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 16(3), 367-378.

Sherman, S. M., & Guillery, R. W. (1998). On the actions that one nerve cell can have on another: distinguishing "drivers" from "modulators". *Proceedings of the National Academy of Sciences*, 95(12), 7121-7126

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587): 484-489.

Smeets JB, Sousa R, Brenner E. 2009 Illusions can warp visual space.Perception 38, 1467
Stich, S & Warfield, T. (Eds.), 1994. *Mental Representation*. Oxford: Blackwell

Snyder, L. H., Grieve, K. L., Brotchie, P., & Andersen, R. A. (1998). Separate body-and world-referenced representations of visual space in parietal cortex. *Nature*, 394(6696), 887-891

Svarverud, E., Gilson, S. and Glennerster, A. (2012). A demonstration of 'broken' visual space. *PloS one*, *7*(3): p.e33782.

Tarr, M. J., & Bülthoff, H. H. (1998). Image-based object recognition in man, monkey and machine. *Cognition*, 67(1), 1-20.

Torr, P. H., & Zisserman, A. (1997). Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing* 15(8): 591-605.

Triggs, B., McLauchlan, P. F., Hartley, R. I., & Fitzgibbon, A. W. (1999). Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms* (pp. 298-372). Berlin / Heidelberg: Springer.

Turi, M., & Burr, D. (2012). Spatiotopic perceptual maps in humans: evidence from motion adaptation. *Proceedings of the Royal Society of London B: Biological Sciences* 279(1740): 3091-3097.

Watt, R. J. (1987). Scanning from coarse to fine spatial scales in the human visual system after the onset of a stimulus. *Journal of the Optical Society of America A* 4(10): 2006-2021.

Waitzman, D. M., Ma, T. P., Optican, L. M., & Wurtz, R. H. (1988). Superior colliculus neurons provide the saccadic motor error signal. *Experimental Brain Research* 72(3): 649-652.

Weiskrantz, L. (1997). Consciousness Lost and Found: A neuropsychological exploration. Oxford: Oxford University Press.

Zhu, Y., Mottaghi, R., Kolve, E., Lim, J. J., Gupta, A., Fei-Fei, L., & Farhadi, A. (2016). Target-driven visual navigation in indoor scenes using deep reinforcement learning. *arXiv preprint arXiv:1609.05143*

Zimmermann, E., Burr, D., & Morrone, M. C. (2011). Spatiotopic visual maps revealed by saccadic adaptation in humans. *Current Biology* 21(16): 1380-1384.

Zipser, D., & Andersen, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature* 331(6158): 679-684.