A moving observer in a three-dimensional world

PHILOSOPHICAL TRANSACTIONS B

Phil. Trans. R. Soc. B.
doi:10.1098/not yet assigned

A moving observer in a 3D world

Andrew Glennerster

School of Psychology and Clinical Language Sciences, University of Reading, UK. OrchidID 0000-0002-8674-2763

Keywords: 3D vision, moving observer, stereopsis, motion parallax, scene representation, visual stability

Summary

For many tasks, such as retrieving a previously viewed object, an observer must form a representation of the world at one location and use it at another. A world-based 3D reconstruction of the scene built up from visual information would fulfil this requirement, something computer vision now achieves with great speed and accuracy. However, I argue that it is neither easy nor necessary for the brain to do this. I discuss biologically plausible alternatives, including the possibility of avoiding 3D coordinate frames such as ego-centric and world-based representations. For example, the distance, slant and local shape of surfaces dictate the propensity of visual features to move in the image with respect to one another as the observer's perspective changes (through movement or binocular viewing). Such propensities can be stored without the need for 3D reference frames. The problem of representing a stable scene in the face of continual head and eye movements is an appropriate starting place for understanding the goal of 3D vision, more so, I argue, than the case of a static binocular observer.

1. Introduction

Many of the papers in this volume consider vision in a 3D world from the perspective of a stationary observer. fMRI, neurophysiological recording and most binocular psychophysical experiments require the participant's head to be restrained. While this can be useful for some purposes, it can also adversely affect the way that neuroscientists think about 3D vision, since it distracts attention from the more general problem that an observer must solve if they are to represent and interact with their environment as they move around. It is logical to tackle the general problem first and then to consider static binocular vision as a limiting case.

Marr famously described the problem of vision as 'knowing what is where by looking' (Marr, 1982). But 'where' is tricky to define. It requires a coordinate frame of some kind and it is not obvious what this (or these) should be. Gibson (1979) emphasised the importance of 'heuristics' by which visual information could be used to control action, such as the folding of a gannet's wings (Lee and Reddish, 1981), without relying on 3D representations and sometimes he appeared to deny the need for representation altogether. However, it is evident that animals plan actions using representations, for example when they retrieve an object that is currently out of view, but the form that these representations take is not yet clear. I will review the approach taken in computer vision, since current systems based on 3D reconstruction work very well, and in the visual system of insects, since they use quite different methods from computer vision systems and yet operate successfully in a 3D environment. Current ideas about 3D representation in the cortex differ in important ways from either of these because biologists hypothesise intermediate representations between image and world-based frames; I will discuss some of the challenges that such models face and describe an alternative type of representation based on quite different assumptions.

2. Possible reference frames

*Author for correspondence (a.glennerster@reading.ac.uk).
†Present address: School of Psychology and Clinical Language Sciences, University of Reading, Reading RG6 6AL, UK

2.1 Computer vision

Computer vision systems are now able to generate a representation of a static scene as the camera moves through it and to track the 6 d.o.f. movement of the camera in the same coordinate frame. This 'Simultaneous Localisation and Mapping' (SLAM) can be done in real time (Davison, 2003) with multiple moving objects (Fitzgibbon and Zisserman, 2000) and even when nothing in the scene is rigid (Fitzgibbon, 2001). Nevertheless, in all these cases the rotation and translation of the camera are recovered in the same 3D frame as the world points. The algorithms are quite unlike those proposed in the cortex and hippocampus since the latter involve a sequence of transformations from eye-centred to head-centred and then world-centred frames (Section 2.3). In computer vision, the scene structure and camera motion over multiple frames are generally recovered in a single step that relies on the assumption of a stable scene (Triggs, McLauchlan, Hartley and Fitzgibbon, 2000).

Early 3D reconstruction algorithms generally identified small, robust features in the input images that can be tracked reliably across multiple frames (e.g. Harris and Stephens, 1988; Fitzgibbon and Zisserman, 1998; Triggs *et al*, 2000). The output is a 'cloud' of 3D points in a world-based frame, each point corresponding to a tracked feature in the input images (e.g. Figure 1a). Modern computer vision systems can carry out this process in real time giving rise to a dense reconstruction (Figure 1b) and a highly reliable recovery of the camera pose (Davison, 2003; Newcombe, Lovegrove and Davison**,** 2011). Many SLAM algorithms now also incorporate an 'appearance-based' element, such as the inclusion of 'keyframes' where the full video frame or omnidirectional view is stored at discrete points along the path which aids re-orientation when normal tracking is lost and helps with 'loop-closure' (Cummins and Newman, 2008; Twinanda, Meilland, Sidibé and Comport, 2013). A 'pose graph' describes the relationship between the keyframes (Figure 1d). Nevertheless, the edges of the graph are 3D rotations and translations and there are local 3D representations at each node. More recent examples abandon the pose graph and demonstrate how it is possible to build a detailed, 3D, global, world-based representation for large-scale movements of the camera (Whelan, Leutenegger, Salas-Moreno, Glocker and Davison, 2015).

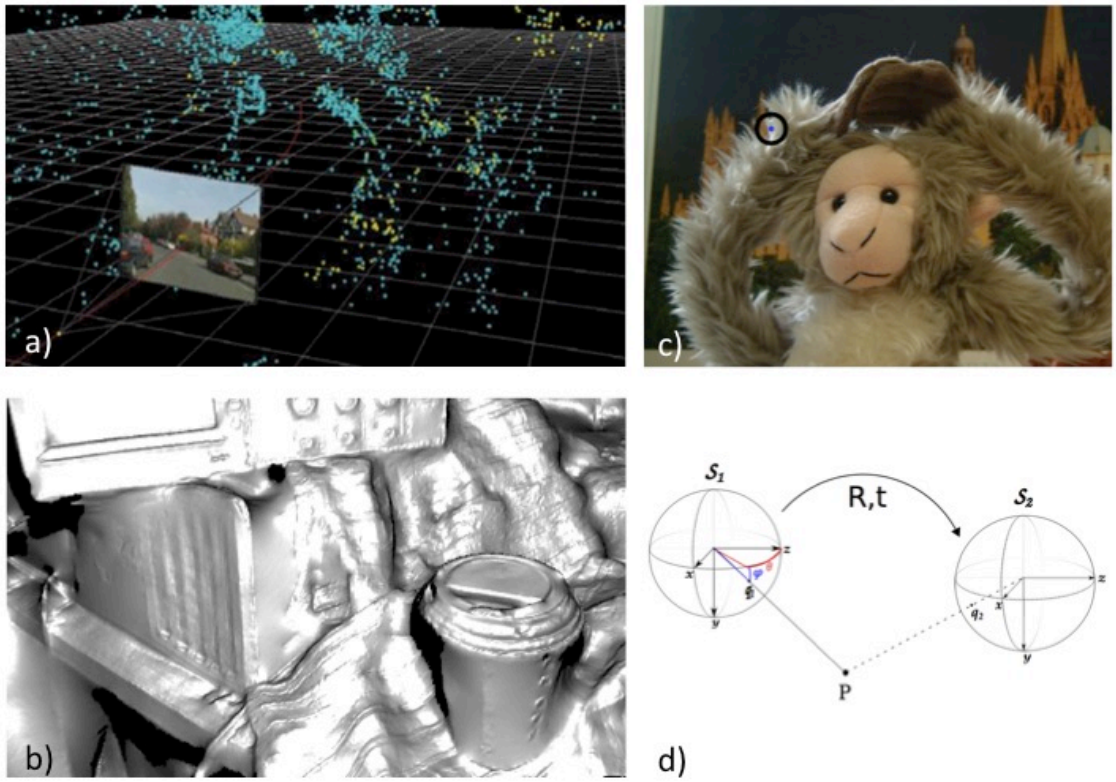

Figure 1. ***Computer vision approaches.*** *a) Early photogrammetry methods tracked features across a sequence of frames and calculated a set of 3D points and a camera path that would best explain the tracks (Image courtesy of Oxford Metrics (OMG plc)). b) 'Dense SLAM' now achieves the same result but for a very dense reconstruction of surfaces and is done in real time (© Reprinted from Newcombe et al, (2011) with permission from IEEE). c) Sometimes it is very difficult to calculate the 3D structure of a scene, as here, and for many purposes solutions that avoid 3D reconstruction are optimal (in this case, synthesising a novel view given several input views, (© Reprinted from Fitzgibbon, Wexler and Zisserman, (2005) with permission from IEEE). d) Recent approaches to SLAM incorporate views at certain locations (S1 and S2*

here, called 'keyframes') and store these, along with the rotation and translation required to move between them, as a graph (Reprinted from Twinanda et al (2013) with permission from the authors).

On the other hand, some computer vision algorithms have abandoned the use of 3D reference frames to carry out tasks that, in the past, might have been tackled by building a 3D model. For example, Rav-Acha, Kohli, Rother and Fitzgibbon (2008) show how it is possible to add a moustache to a video of a moving face captured with a hand-held camera. In theory, this could be achieved by generating a deformable 3D model of the head, but the authors' solution was to extract a stable texture from the images of the face (an 'unwrap mosaic'), add the moustache to that and then 'paste' the new texture back onto the original frames. The result appears convincingly '3D' despite the fact that no 3D coordinates were computed at any stage. Closely related image-based approaches have been used for a localisation task (Ni, Kannan, Criminisi and Winn, 2009). In the movie industry and in many other applications, the start and end point are images, in which case an intermediate representation in a 3D frame can often be avoided. Another case is image interpolation, using images from two or more cameras. In theory, this can be done by computing the 3D structure of the scene and projecting points back into a new, simulated camera. But for some objects, like the fluffy toy in Figure 1c, this is hard. The best-looking results are obtained by instead optimising for 'likely' image statistics in the simulated scene, using the input frames to determine these statistics (Fitzgibbon et al, 2005) and once again avoiding the generation of a 3D model. A very similar argument can be applied to biology. Both the context for and the consequence of a movement are a set of sensory signals (Miall, Weir, Wolpert and Stein, 1993), so it is worth considering whether the logic developed in computer graphics might also be true in the brain, i.e. that a non-3D representation might do just as well (under most conditions) as a putative internal 3D model.

2.2 Image-based strategies

It is widely accepted that animals achieve many tasks using 'image-based' strategies, where this usually refers to the control of some action by monitoring a small number of parameters such as the angular size of an object on the retina and/or its rate of expansion as the animal moves. Even in insects, these strategies can be quite sophisticated. Cartwright and Collett (1983) showed how bees remember and match the angles between landmarks to find a feeding site and, when the size of a landmark is changed, they alter their distance to match the retinal angle with the learned size. Ants show similar image-based strategies in returning to a place or following a route (Wehner and Raber, 1979; Graham and Collett, 2002). Equally, it is widely accepted that many simple activities in humans are probably achieved using image-based rules, such as the online correction of errors in reaching movements (Lawrence, Khan, Buckolz, & Oldham, 2006; Sarlegna, Blouin, Vercher, Bresciani, Bourdin and Gauthier, 2004), and the fixation locations chosen by the visual system during daily activities often make it particularly easy for the brain to monitor visual parameters that are useful for guiding action, e.g. fixating a target object and bringing the image of the hand towards the fovea (Land, Mennie and Rusted, 1999). There is evidence for many tasks being carried out using simple strategies including cornering at a bend (Wilkie, Wann and Allison, 2008), catching a fly-ball (McBeath, Shaffer and Kaiser, 1995) or timing pull shot in cricket (Land and McLeod, 2000) and there is a long history of using 2D image-based strategies to control robots (Braitenberg, 1986).

Movements take the observer from one state to another (hand position, head position, etc) and hence, in the case of visually guided movements, from one set of image-based cues (or sensory context) to another. Image-based strategies are *ad hoc,* unlike a 'cognitive map' whose whole purpose is to be a common resource available to guide many different movements (O'Keefe and Nadel, 1978). But image-based strategies require *some* sort of representation that goes beyond the current image. Gillner and Mallot (1998), for example, have measured the ability of participants to learn the layout of a virtual town, navigate back to objects and find novel shortcuts. They suggested that people's behaviour was consistent with them building up a 'graph of views', where the edges are actions (forward movement and turns) and the nodes are views (Figure 2b). Similarly, data by Schnapp and Warren (in abstract form, Schnapp and Warren (2007) and Warren, Rothman, Schnapp and Ericson (submitted)) have tested participants' ability to navigate in a VR environment that does not correspond to any possible metric structure. It contains 'wormholes' that transport participants to a different location and different orientation in the maze without them being aware that this has happened (see Figure 2a). Because they are translated and rotated as they move through the wormhole, no coherent (wormhole-free) 2D map of the environment is possible. The fact that participants do not notice and can perform navigation tasks well suggests that they are not using a cognitive map but their behaviour is consistent with a topological representation formed of a graph of views.

Some behaviours are more difficult to explain within a view-based or sensory-based framework. Path-integration by ants provides a good example: they behave as if they have access to a continually updated vector (direction and distance) that will take them back to home. This can be demonstrated by transporting an ant that has walked away from its nest and releasing it at a new location (Müller and Wehner, 1988). Müller

and Wehner do not use this evidence to propose a cognitive map in ants but instead argue that the pattern of systematic errors in the path integration process suggests that the ants are using a simple mechanism that is closely analogous to homing by matching a 'snapshot', i.e. a classic image-based strategy. Nevertheless, for some tasks it is difficult to imagine image-based strategies: 'Point to Paris!', for example. People are not always good at these tasks and they often require considerable cognitive effort. It is not yet clear whether, in these difficult cases, the brain resorts to building a 3D reconstruction of the scene or whether there are yet-to-be-determined view-based approaches that could account for them.

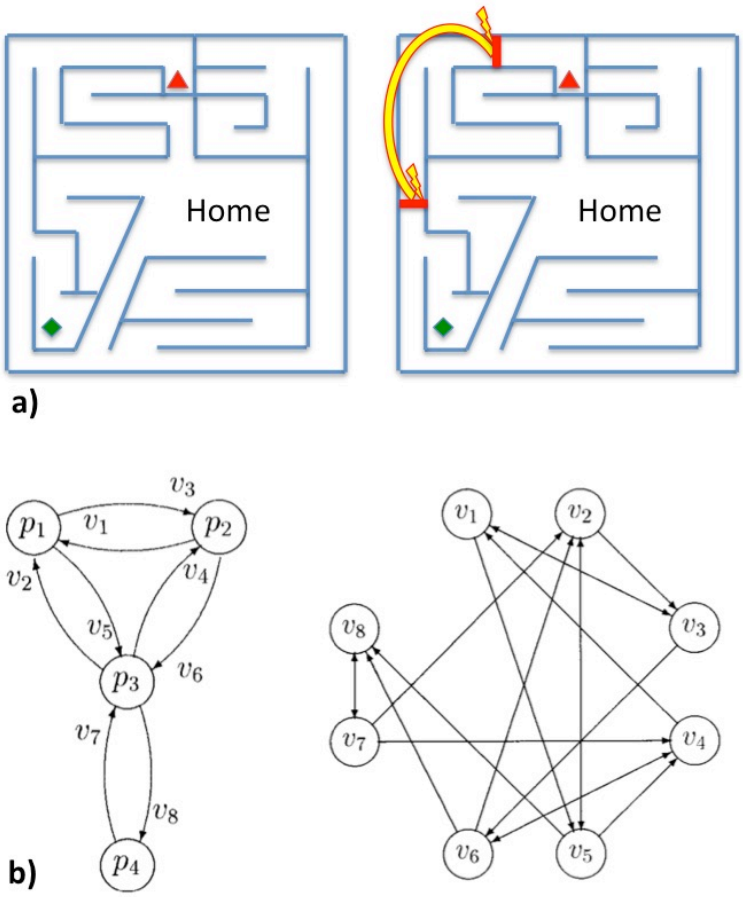

Figure 2. ***Non-Euclidean representations.*** *a) In an experiment by Schnapp and Warren (2007), participants explored a virtual environment that either corresponded to a fixed Euclidean structure (left) or something that was not Euclidean (right) because, in this case, participants were transported through a 'wormhole' between the locations marked on the map by red lines but the views from these two locations were identical so there was no way to detect the moment of transportation. In the wormhole condition, the relative location of objects has no consistent geometric interpretation (sketch of virtual maze adapted from Schnapp and Warren (2007)). b) Four places, p1-p4, are shown as nodes in a graph whose edges are the views from each place. The views themselves can be described as a graph (right). In this case, the edges are actions (rotations on the spot or translations between places).*

2.3 Cortical representations

It is often said that posterior parietal cortex represents the scene in a variety of coordinate frames (Andersen, Snyder, Bradley and Xing, 1997; Colby, 1998), as shown in Figure 3a. The clearest case for something akin to a 3D coordinate frame is in V1. Here, receptive fields are organised retinotopically and neurons are sensitive to a range of disparities at each retinal location (e.g. Prince, Cumming and Parker, 2002). Described in relation to the scene, this amounts (more or less) to a 3D coordinate frame centred on the fixated object. In theory, a rigid

rotation and translation could transform the 3D receptive fields in V1 into a different frame, e.g. one with an origin and axes attached to the observer's hand. If this were the case, one would expect a very rapid and quite complex re-organisation of receptive fields in posterior parietal cortex as the hand translated or rotated. But that would be substantially more complex than the type of operations that have been proposed up to now. For example, 'gain fields' demonstrate that one parameter, such as the position of the eyes in the head, can modulate the response of neurons to visual input (Zipser and Andersen, 1988; Buneo and Andersen, 2006). In some cases, operations of this type can give rise to receptive fields that are stable in a non-retinotopic frame (e.g. Duhamel, Bremmer, BenHamed and Graf, 1997).

Beyond posterior parietal cortex, a further coordinate transformation is assumed to take place to bring visual information into a world-based frame. Byrne, Becker and Burgess (2007) describe steps that would be required to achieve this (Figure 3b). An ego-centred representation, assumed to come from parietal cortex, would need to be duplicated many times over so that a signal from a head direction cell could 'gate' the information passing to 'Boundary Vector Cells' (BVC) in parahippocampal gyrus. This putative mechanism deals with rotation. A similar duplication would presumably be required to deal with translation.

Anatomically nearby, but quite different in their properties, 'grid cells' in the dorsocaudal medial entorhinal cortex provide information about a rat's location as it moves (Hafting, Fyhn, Molden, Moser and Moser, 2005). The signals from any one of these neurons are highly ambiguous about the rat's location. 3 grid cells at each of 3 spatial scales could, in theory, signal a very large number of locations, just as 9 digits can be used to signal a million different values, but the readout of these values to provide an unambiguous signal that identifies a large number of different locations would be difficult (Bush, Barry, Manson and Burgess, 2015) especially if realistic levels of noise in the grid cells were to be modelled. Instead, arguments have been advanced that place cell receptive fields are not built up from grid cell input but that, instead, information from 'place' and 'grid' cells complement one another (Bush, Barry and Burgess, 2014; Poucet, Sargolini, Song, Hangya, Fox, and Muller, 2014). In relation to this volume, which is about vision in a 3D world, it is relevant to note that grid cells are able to operate very similarly in the dark and the light (Hafting *et al*, 2005), so the visual coordinate transformations discussed above are clearly not necessary to stimulate grid cell responses. Indeed, some have argued that grid cells play a key role in navigation only in the dark (Poucet *et al*, 2014).

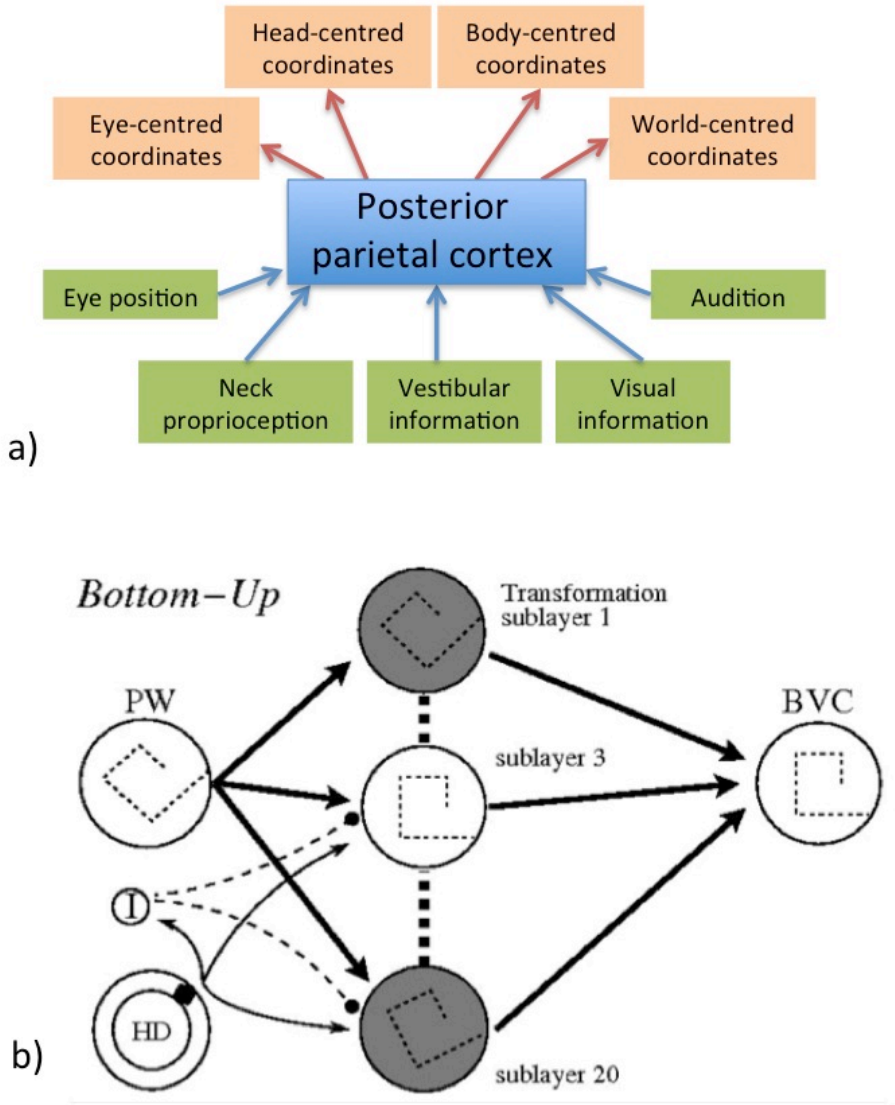

Figure 3. ***Putative neural representations of 3D space****. a) This diagram, adapted from Andersen* et al *(1997), reflects a common assumption that parietal cortex in primates transforms sensory information of different types into 3D representations of the scene in a variety of different ego-centric coordinate frames. b) Byrne* et al *(2007) propose a mechanism for transforming an ego-centric representation into a world-based one using the output of head-direction cells. A set of identical populations of neurons (20 in this example, nominally in the retrosplenial cortex) each encode a repeated version of the scene but rotated by different amounts, on the basis of an ego-centric input from parietal cortex (PW). A signal from head-direction (HD) cells could 'gate' the information and so ensure that the output to Boundary Vector Cells (BVC), which are hypothesised to exist in parahippocampal cortex, is maintained in a world-centred frame.*

2.4 Removing the origin

Instead of representing space as a set of receptive fields, e.g. with 3D coordinates defined by visual direction and disparity in V1 as discussed in Section 2.3, an alternative is that it could be represented by a set of sensory contexts and the movements that connect these (like a graph, as discussed in section 2.1 and 2.2). Similarly, shape and slant could be represented by storing the propensity of a shape to deform in particular ways as the observer moves - again, it is the linking of sensory contexts via motor outputs that is important (Glennerster, Hansard and Fitzgibbon, 2009). The idea is not to store all motion parallax as an observer moves. Any system that did this would be able (in theory) to compute the 3D structure of the scene and the trajectory of the

observer. Instead, the idea is that what is stored is some kind of 'summary' that is useful for action despite being incomplete. I will refer throughout this section to *movement* of the observer (or optic centre), but the key is that the scene is viewed from a number of different vantage points. A limiting case is just two vantage points, including a static binocular observer, but the principles should apply to a much wider range of eye and head movements.

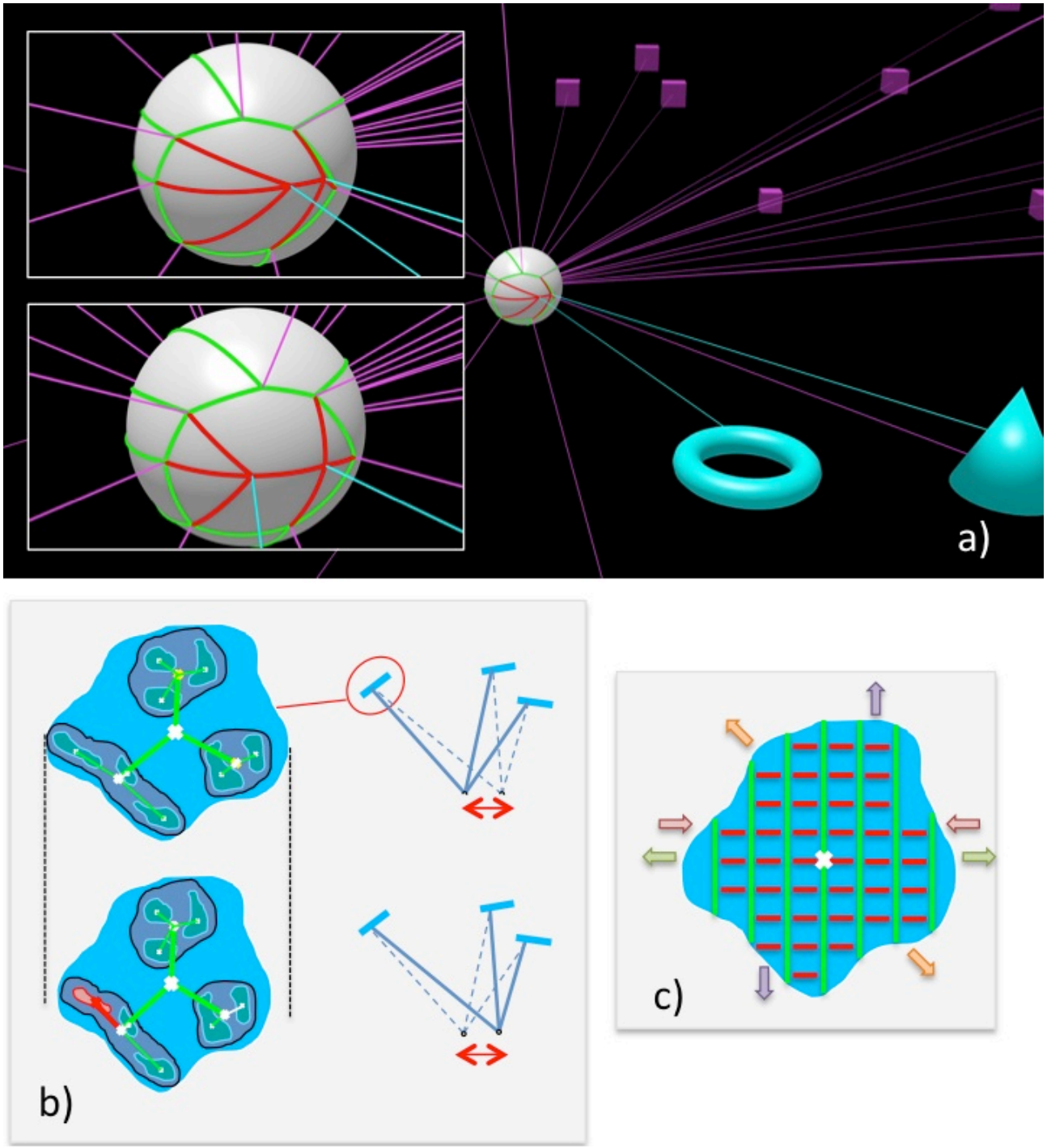

Figure 4. ***Image consequences of observer movement****. Information about the distance, slant and local shape of surfaces can be gathered from the propensity of features to move relative to one another in the optic array as the observer moves in different directions in a static scene. a) Static distant objects (in this case, purple cubes) do not change their relative visual direction as the observer moves (green arcs) whereas the image of near objects (shown in cyan) change relative to distant objects and relative to each other (red arcs). The white sphere shows the optic array around an optic centre in the centre of the sphere. The inset images show the optic array for two different locations of the optic centre (same static scene). Purple rays come from distant cubes, cyan rays come from near, cyan objects. The lengths of the arcs and the*

angles between them provide a description of the 'relative visual direction' of features in the optic array. For the green arcs, these remain stable despite reasonably large translations of the optic centre. b) Considering a much smaller region of the optic array, a patch on a surface can also be considered in terms of features that remain constant despite observer movement versus those that change. The patch shown is slanted with respect to the line of sight and compresses horizontally as the observer translates. Most of the features compress in the same way as the overall compression of the patch (green arcs), as if drawn on a rubber sheet, whereas one feature moves (shown in red). It has depth relief relative to the plane of the surface patch (Glennerster and McKee, 2004). c) The tilt and qualitative information about the slant of the patch are indicated by the distortion of the rubber sheet. For any component of observer translation in a plane orthogonal to the line of sight (i.e. not approaching or backing away from the surface), different directions of observer movement produce effects shown by the coloured arrows (compression - red; expansion - green; shear - purple; a mixture - orange). The green rods stay the same length and remain parallel throughout; their orientation defines the tilt of the surface. The red lines joining the greed rods indicate the 'elasticity' of the rubber sheet: the more elastic they are the greater the surface slant. Any component of observer translation towards the surface causes a uniform expansion of the whole sheet (including the green rods).

For example, consider a camera or eye rotating about its optic centre so that, over many rotations, it can view the entire panoramic scene or so-called 'optic array'. If the axis and angle that will take the eye/camera from any point on this sphere to any other is recorded, then these relative visual directions provide a framework to describe the layout or 'position' of features across the optic array (Glennerster, Hansard and Fitzgibbon, 2001; Figure 4a). In practice, the number of relative visual directions that need to be stored can be significantly reduced by organising features hierarchically, e.g. storing finer scale features within coarser scale ones (Watt and Morgan, 1985; Watt, 1987, Glennerster *et al*, 2001, see Figure 4b). This means that every fine scale feature has a location within a hierarchical database of relative visual directions. Saccades allow the observer to 'paint' extra detail into the representation in different regions, like a painter adding brush strokes on a canvas, but the framework of the canvas remains the same (Bridgeman, Van der Heijden and Velichkovsky, 1994).

When the optic centre translates (including the case of binocular vision, where the translation is from one eye to the other), information becomes available about the distance, slant and depth relief of surfaces:

- *Distance:* Some features in the optic array remain relatively stable with respect to each other when the optic centre translates (Glennerster *et al*, 2001). Examples of this are shown in Fig 4a (shown in green). For large angular separations, when pairs or triples of points do not move relative to one another in the face of optic centre translation, the points must be distant. These points form a stable background against which the parallax (or disparity) of closer features can be judged (shown in red in Fig 4a). This type of representation of 'planes plus parallax' is familiar in computer vision, where explicit recovery of 3D structure can be avoided, and many tasks simplified, by considering a set of points in a plane (sometimes these are points at infinity) and recording parallax relative to these points (Anandan, Irani, Kumar and Bergen, 1995; Triggs, 2000; Torr, Szeliski and Anandan, 2001). Representing the propensity of features to move relative to a stable background allows one to encode information about the relative distance of objects without necessarily forming a 3D, world-based representation.

- *Depth relief:* Given that the definition of visual direction of features is recorded hierarchically in the proposed representation, there is a good argument for storing deformations in a hierarchical way, too. So, if a surface is slanted and translation of the optic centre causes a lateral compression of the image then the basis vector or coordinate frame for recording the visual direction of finer scale features should become compressed too. Koenderink and van Doorn (1991) describe the advantages of using a 'rubber sheet' coordinate system like this. It has the effect that features on the slanted plane are recorded as having 'zero' disparity (or motion) and any disparity (or motion) signals a 'bump' on the surface (shown in red in Fig 4b). There is good evidence that the visual system adopts a 'rubber sheet' coordinate frame of this sort from experiments on binocular correspondence (Mitchison and McKee,1987), perceived depth (Mitchison and Westheimer, 1984) and stereoacuity (Glennerster, McKee and Birch, 2002; Petrov and Glennerster, 2006).

- *Slant:* Fig 4c shows how the slant of a surface patch might be represented in such a way that is short of a full metric 3D description of its angle and yet useful for many purposes. Information about the image deformation of a surface patch provides information about the angle of tilt of the surface and some qualitative information about the magnitude of its slant (e.g. Koenderink, 1986). Figure 4c shows how moving in different directions is, in image terms, a bit like pulling and pushing a rubber sheet that contains rigid parallel rods: the more slanted the surface, the more elastic the connection between the green rods (shown as red lines in Fig 4c). The tilt of the rods indicates the tilt of the surface. The rods can change in length (and the whole patch with them) if the optic centre moves towards or away from the surface.

Thus, for distance, slant and depth relief we have identified information about the propensity of 2D image quantities to change in response to observer movement (or binocular viewing). It is important to emphasise that this 'propensity' to change is neither optic flow (Koenderink, 1986; Bruhn, Weickert and Schnörr, 2005) nor a 3D reconstruction but somewhere in between. Figure 5 illustrates the point. Viewing a slanted surface gives rise to different optic flow depending on the direction of translation of the optic centre (Fig 5a) but if the visual system is to use the optic flow generated by one translation to predict the flow that will be generated by a different head movement then it must infer something general about the surface. This need not be the 3D structure of the surface, although clearly that would be one general description that would support predictions. Instead, the visual system might store something more image-based, as illustrated in Figure 4. Neurally, this could be instantiated as a graph of sensory states joined by actions (Glennerster *et al*, 2009; Marr, 1969; Albus, 1971). For example, if all the images shown in Fig 5a correspond to nodes in a graph (Fig 5c) and translations of the optic centre corresponded the edges then the relationship between the nodes and the edges carries information about the surface slant: if a relatively large translation is required to move between nodes (the 'propensity' of the image to change with head translation is relatively low), then the surface slant is shallow. The same type of graph could underlie the idea of a 'canvas' on which to 'paint' visual information as the eyes move (Fig 5b), as discussed above. The edges in this case are saccades (MacKay, 1973; O'Regan and Noë, 2001; Glennerster, 2015; Glennerster, Hansard and Fitzgibbon, 2001, 2009).

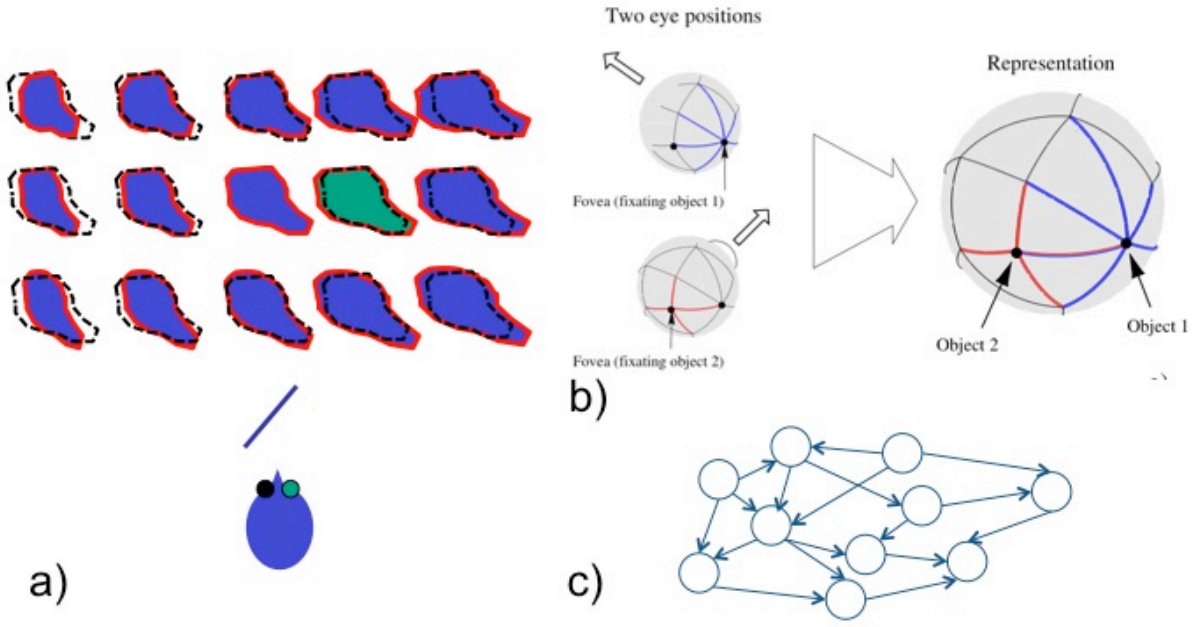

Figure 5. ***What might be stored****? a) An observer views a slanted surface, as shown in plan view. The image as seen from the left eye is shown in the centre (red outline). The image immediately to the right (in green) shows how the right eye receives an image that is expanded laterally. The black dashed line shows the outline of the surface in the original (central) image. The top and bottom rows show the image consequences of the original (left eye) viewpoint moving up or down respectively and the columns show the consequences of the viewpoint moving left or right. These transformations can be understood in relation to a stable 3D structure of the surface or in terms of the 'propensity' of parts of the image to change when the viewpoint moves. b) An eye is shown viewing a static scene and rotating about its optic centre. The image from one viewing direction (shown in blue) overlaps with the image from another viewing direction (shown in red) and both these images can be understood in relation to a stable 2D sphere of visual directions, as shown on the right. Reprinted from Glennerster* et al *(2001) with permission from Elsevier. c) As discussed in the text, both these concepts ('propensity' and a stable 'canvas' on which to paint images) might be implemented using a graph where the nodes are sensory+motivational states and the edges are actions (translations of the head in a) or rotations of the eye in b)).*

Taken together, we now have a representation with many of the properties that Marr and Nishihara (1978) proposed when they discussed a 2½-D sketch. The eye can rotate freely and the representation is unchanged. Small translations in different directions (including from left to right eye, i.e. binocular disparity) give information about the relative depth of a surface, its slant relative to the line of sight and the relief of fine scale features relative to the plane of the surface. It is, in a sense, an ego-centric representation in that the eye is at the centre of the sphere. Yet, in another sense, it is world-based, since distant points remain fixed in the

representation, independent of the rotation and translation of the observer. This applies to an observer in a scene, moving their head and eyes, which is the situation Marr and Nishihara envisaged when they described their 2½-D sketch. For larger translations, such as walking out of the room, a graph of views is still an appropriate representation (Gillner and Mallot, 1998) but many of the relationships illustrated in figure 4 would no longer apply for such 'long baseline' translations.

The purpose of Marr's primal sketch (Marr and Hildreth, 1980) and the 2½-D sketch was that they were summaries, where information was made explicit if it was useful to the observer. That is also a feature of the representation described here, in that information can be left in 'summary form' or filled in in greater detail when required. In the case of an eye/camera rotating about its optic centre, there is 'room' in the representation to 'paste in' as much fine scale detail as is available to the visual system (Glennerster, 2013), even though, under most circumstances, observers are unlikely to need to do this and, as many have argued, fine scale detail need not be stored in a representation if it can be accessed readily by a saccadic eye movement when required ('assuaging epistemic hunger' as soon as it arises (Dennett, 1993)).

Similarly, there is nothing to stop the visual system using the information from disparity or motion in the representation in a more sophisticated and calibrated way than simply recording measures such as 'elasticity' between features as outlined above. For example, it has been proposed that there is a hierarchy of tasks using disparity information (Tittle, Todd, Perotti and Norman, 1995; Glennerster, Rogers and Bradshaw, 1996) which goes from breaking camouflage at the simplest level through threading a needle, determining the bas-relief structure of a surface, comparing the relief of two surfaces at different distances, and, at the top of the hierarchy, judging the Euclidean shape of a surface. These tasks lie on a spectrum in which more and more precise information is required, either about the disparities produced by a surface or about its distance from the observer, in order to carry out a task successfully. Judgement of Euclidean shape demands a calibration of disparity information – a precise delineation of the current sensory context relative to others - to such an extent that performance is compatible with the brain generating a full Euclidean representation. What is important in this way of thinking, though, is that the top level of calibration of disparities is only carried when the task demands it (which is likely to be rare and, in the example in Section 3 below, occurs only once). If the visual system never used 'summaries' or short-cuts, then the storage of disparity and motion information would be equivalent to a full metric model of the environment.

3. Examples

3.1. *An example task.*

As we said at the outset, for a representation to be useful to a moving observer, information gained in one location must be capable of guiding action in another. Consider a task: a person has to retrieve a mug from the kitchen, starting from the dining room. How can this be achieved, unless the brain stores a 3D model of the scene? The first step is to rotate appropriately, which requires two things. The visual system must somehow know that the mug is behind the one door rather than another, even if it does not store the 3D location of the mug. In computer graphics applications, 'portals' are used in a related way to upload detailed information about certain zones of the virtual environment only when required (Eberly, 2006). The problem of knowing whether an item is down one branch or another of a deep nested tree structure is well known in relational databases (Kothuri, Ravada, Sharma and Banerjee, 2003) but is not considered here. Second, the direction and angle of the kitchen door relative to the current fixated object must be stored in the representation since it is currently out of view. Next, the person must pass through the door and rotate again so as to bring the mug into view. Each of these states and transitions could be considered as nodes and edges in a graph, with fairly simple actions joining the states.

The final parts of the task require a different type of interaction with the world, since the person must reach out and grasp the mug. The finger and thumb must be separated by an appropriate amount in order to grasp the mug and there is good evidence that information about the metric shape of an object begins to affect the grasp aperture before the hand comes into view (Jeannerod, 1984; Servos, Goodale, & Jakobson, 1992, Watt and Bradshaw, 2003). It is possible that information from a number of sources, both retinal and extra-retinal, about the distance and metric structure of the target object can be brought to bear just at this moment (because, once the hand is in view, closed loop visual guidance can play a part (Watt and Bradshaw, 2003; Makin, Holmes, Brozzoli and Farnè, 2012, Sarlegna *et al.*, 2004; Saunders & Knill, 2004)). As discussed above, the fact that a range of sources of relevant information can be brought to bear at a critical moment (when shaping the hand before a grasp) makes it very difficult to devise an experimental test that could distinguish between the predictions of competing models, i.e. a graph-based representation versus one that assumes the mug, hand and observer's head are all represented in a common 3D reference frame with the shape of the mug described in full, Euclidean, metric coordinates.

That may seem slippery, from the perspective of designing critical experiments, but it is an important point. A graph of contexts connected by actions is a powerful and flexible notion. If the task is only to discriminate between bumps and dips on a surface, then the contexts that need to be distinguished can be very broad ones (e.g. positive versus negative disparities). On the other hand, the context corresponding to a mug with a particular size, shape and distance is a lot more specific. It may require not only more specific visual information but also information from other senses such as proprioception, including vergence, in order to narrow it down. All the same, it remains a context for action. During the mug-retrieving task, most of the steps do not require such a narrow, richly-defined context including all the stages at which visual guidance is possible or the action is a pure rotation of the eye and head. But some do, and in those cases it is possible to specify more precise, multi-sensory contexts that will discriminate between different actions.

3.2. Example predictions

When putting forward their stereo algorithm, Marr and Poggio (1979) went to admirable lengths to list psychophysical and neurophysiological results that, if they could be demonstrated, would falsify their hypothesis. Here are a few that would make the proposals described above untenable (Section 2.4). Using Marr and Poggio's, convention, the number of stars by a prediction (*'P'*) indicates the extent to which the result would be fatal and *'A'* indicates supportive data that already exist.

- *(P***). Coordinate transformations.* Strong evidence in favour of true coordinate transformations of visual information in the parietal cortex or hippocampus would be highly problematic for the ideas set out in Section 2.4. If it could be shown that visual information in retinotopic visual areas like V1 goes through a rotation and translation *'en masse'* to generate receptive fields with a new origin and rotated axes in another visual area, where these new receptive fields relate to the orientation of the head, hand or body then the ideas set out in Section 2.4 will be proved wrong, since they are based on a quite different principle. Equally fatal would be a demonstration that the proposal illustrated in Figure 3b is correct, or any similar proposal involving multiple duplications of a representation in one coordinate frame in order to choose one of the set based on idiothetic information. Current models of coordinate transformations in parietal cortex are much more modest, simulating 'partially shifting receptive fields' (Pouget, Deneve and Duhamel, 2002) or 'gain fields' (Zipser and Andersen, 1988) which are 2D, not 3D transformations. Similarly, models of grid cell or hippocampal place cell firing do not describe how 3D transformations could take place taking input from visual receptive fields in V1 and transforming them into a different, world-based 3D coordinate frame (Burgess and O'Keefe, 1996; Burgess, 2008; Whitlock, Sutherland, Witter, Moser and Moser, 2008).

- *(P***). World-centred visual receptive fields.* This does not refer to receptive fields of neurons that respond to the location of the observer (O'Keefe, 1979). After all, the location of the observer is not represented by a V1 receptive field (it is invisible) so no rotation and translation of visual receptive fields from retinotopic to egocentric to world-centred coordinates could make a place cell. A world-centred visual receptive field is a 3D 'voxel' much like the 3D receptive field of a disparity-tuned neuron in V1 but based in world-centred coordinates. Its structure is independent of the test object brought into the receptive field and independent of the location of the observer or the fixation point. For example, if the animal viewed a scene from the South and then moved, in the dark, round to the West, evidence of 3D receptive fields remaining constant in a world-based frame would be incompatible with the ideas set out here. In this example, the last visual voxels to be filled before the lights went out should remain in the same 3D location, contain the same visual information (give or take general memory decay across all voxels) and remain at the same resolution, despite the translation, rotation and new fixation point of the animal. An experiment that followed this type of logic but for pointing direction found, on the contrary, evidence for gaze-centred encoding (Henriques, Klier, Smith, Lowy and Crawford, 1998).

- *(A*) Task-dependent performance.* If all tasks are carried out with reference to an internal model of the world (a 'cognitive map' or reconstruction), then whatever distortions there are in that model with respect to ground truth should be reflected in all tasks that depend on that model. Proof that this is the case would make the hypothesis set out in Section 2.4 untenable. However, there is already considerable evidence that the internal representation used by the visual system is something much looser and, instead, that different strategies are used in response to different tasks. Many examples demonstrate such 'task-dependence' (Koenderink, van Doorn, Kappers and Lappin, 2002; Smeets, Sousa and Brenner, 2009; Svarverud, Gilson and Glennerster, 2012; Glennerster *et al*, 1996; Knill, Bondada and Chhabra, 2011). For example, when participants compare the depth relief of two disparity-defined surfaces at different distances they do so very accurately while, at the same time, having substantial biases in depth-to-height shape judgements (Glennerster *et al*, 1996). This

experiment was designed to ensure that, to all intents and purposes, the binocular images the participant received were the same for both tasks so that any effect on responses was not due to differences in the information available to the visual system. The fact that biases were systematically different in the two tasks rules out the possibility that participants were making both judgements by referring to the same internal 'model' of the scene. Discussing a related experiment that demonstrates inconsistency between performance on two spatial tasks, Koenderink *et al* (2002) suggest that it might be time to "… discard the notion of `visual space' altogether. We consider this an entirely reasonable direction to explore, and perhaps in the long run the only viable option."

- *(P**) Head-centred adaptation.* A psychophysical approach could be, for example, to look for evidence of receptive fields that are constant in head-centred coordinates. For example, if an observer fixates a point 20 degrees to the right of the head-centric midline and adapts to a moving stimulus 20 degrees to the left of fixation (i.e. on the head-centric midline), do they show adaptation effects in a head-centric frame after they rotate their head to a new orientation while maintaining fixation (see Figure 6)? Evidence of a pattern of adaptation that followed the head in this situation would not be expected according to the ideas set out in Section 2.4. As Figure 6 illustrates, this prediction is different from either retinal or spatiotopic (world-based) adaptation (Melcher, 2005; Knapen, Rolfs and Cavanagh, 2009; Turi and Burr, 2012). There is psychophysical evidence that gaze direction can modulate adaptation (Mayhew, 1973; Nishida, Motoyoshi, Andersen and Shimojo, 2003) consistent with physiological evidence of 'gain fields' in parietal cortex (Zipser and Andersen, 1988) but the data do not show that adaptation is spatially localised in a head-centred frame as illustrated in Figure 6.

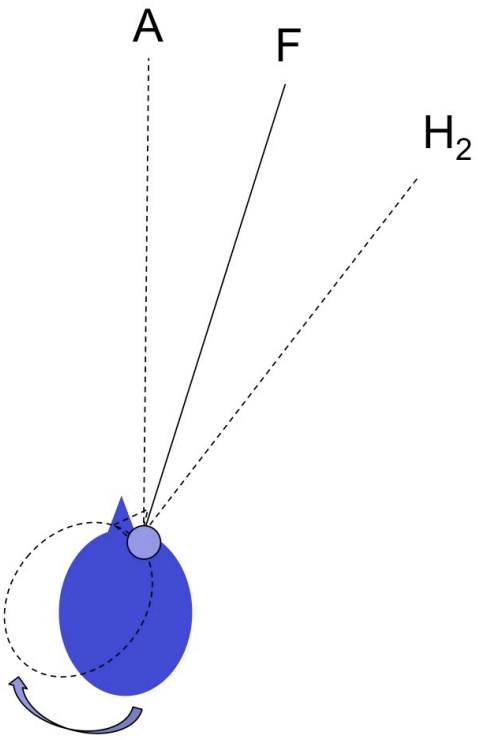

Figure 6. ***Potential evidence of head-centric adaptation.*** *If a participant were to fixate a point 'F' and adapt to a stimulus (e.g. a drifting grating) presented at 'A' in the head-centric midline then turn their head to point in the direction 'H_2' and test the effect of adaptation over a wide range of test directions, the predictions of retinotopic, spatiotopic and head-centric adaptation would differ. In this case, the peak effects should occur at 'A' for retinotopic and spatiotopic adaptation and at 'H_2' for head-centric adaptation.*

4. Conclusion

If a moving observer is to use visual information to guide their actions, then they need a visual representation that encodes the spatial layout of objects. This need not be 3-dimensional, but it must be capable of representing the current state, the desired state and the path between the two. Computer vision

representations are 3D, predominantly, as are most representations that are hypothesised in primates but, I have argued, there is good reason to look for alternative types of representation that avoid 3D coordinate frames.

Acknowledgments
I am grateful for comments from Suzanne McKee , Michael Milford, Jenny Read and Peter Scarfe.

Competing Interests
'I have no competing interests.'

Funding
Funded by EPSRC EP/K011766/1 and EP/N019423/1.

References

2d3 Ltd. Boujou 2, 2003. http://www.2d3.com

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

Anandan, P., Irani, M., Kumar, R., & Bergen, J. (1995). Video as an image data source: efficient representations and applications. In *Image Processing, 1995. Proceedings., International Conference on* (Vol. 1, pp. 318-321). IEEE.

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.

Braitenberg, V. (1986). *Vehicles: Experiments in Synthetic Psychology*. MIT Press

Bridgeman, B., Van der Heijden, A. H. C., & Velichkovsky, B. M. (1994). A theory of visual stability across saccadic eye movements. *Behavioral and Brain Sciences*, *17*(2), 247-257.

Bruhn, A., Weickert, J., & Schnörr, C. (2005). Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, *61*(3), 211-231.

Burgess, N. (2008). Grid cells and theta as oscillatory interference: theory and predictions. Hippocampus, 18(12), 1157-1174

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

Bush, D., Barry, C., & Burgess, N. (2014). What do grid cells contribute to place cell firing?. *Trends in Neurosciences*, *37*(3), 136-145

Bush, D., Barry, C., Manson, D., & Burgess, N. (2015). Using grid cells for navigation. *Neuron*, *87*(3), 507-520.

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), 340-375

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

Cartwright, B. A., & Collett, T. S. (1983). Landmark learning in bees. *Journal of Comparative Physiology*, *151*(4), 521-543.

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

Cummins, M., & 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 *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on* (pp. 1403-1410). IEEE.

Dennett, D. C. (1993). *Consciousness explained*. Penguin UK.

Duhamel, J. R., Bremmer, F., BenHamed, S., & Graf, W. (1997). Spatial invariance of visual receptive fields in parietal cortex neurons. *Nature*, 389(6653), 845-848.

Eberly, D. H. (2006). *3D Game Engine Design: A Practical Approach to Real-time Computer Graphics*. CRC Press.

Fitzgibbon, A. W. (2001). Stochastic rigidity: Image registration for nowhere-static scenes. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on* (Vol. 1, pp. 662-669). IEEE.

Fitzgibbon, A., Wexler, Y., & Zisserman, A. (2005). Image-based rendering using image-based priors. *International Journal of Computer Vision*, 63(2), 141-151.

Fitzgibbon, A. W. and Zisserman, A. (1998) Automatic camera recovery for closed or open image sequences. In *Lecture Notes in Computer Vision* 1406: Computer Vision—ECCV '98, p311–326. Springer,

Fitzgibbon, A. W., & Zisserman, A. (2000). Multibody structure and motion: 3-D reconstruction of independently moving objects. In *Computer Vision-ECCV 2000* (pp. 891-906). Springer Berlin Heidelberg.

Gibson, J.J (1979) *The ecological approach to visual perception*. Boston: Houghton Mifflin

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. (2013) Representing 3D shape and location. In Dickinson, S. and Pizlo, Z. (Ed) *Shape perception in human and computer vision: an interdisciplinary perspective*, Springer-Verlag, London

Glennerster, A. (2015) Visual stability - what is the problem? *Frontiers in Psycholology,* 6:958. doi: 10.3389/fpsyg.2015.00958

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., Hansard, M. E., & Fitzgibbon, A. W. (2009). View-based approaches to spatial representation in human vision. In *Statistical and Geometrical Approaches to Visual Motion Analysis* (pp. 193-208). Springer Berlin Heidelberg.

Glennerster, A., & McKee, S. (2004). Sensitivity to depth relief on slanted surfaces. *Journal of Vision*, 4(5), 3.

Glennerster, A., McKee, S. P., & Birch, M. D. (2002). Evidence for surface-based processing of binocular disparity. *Current Biology*, *12*(10), 825-828.

Glennerster, A., Rogers, B. J., & Bradshaw, M. F. (1996). Stereoscopic depth constancy depends on the subject's task. *Vision Research*, *36*(21), 3441-3456.

Graham, P., & Collett, T. S. (2002). View-based navigation in insects: how wood ants (Formica rufa L.) look at and are guided by extended landmarks. *Journal of Experimental Biology*, *205*(16), 2499-2509.

Hafting, T., Fyhn, M., Molden, S., Moser, M. B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801-806.

Harris, C and Stephens, M (1988). A combined corner and edge detector *Proceedings of the 4th Alvey Vision Conference.* pp. 147–151.

Henriques, D. Y., Klier, E. M., Smith, M. A., Lowy, D., & Crawford, J. D. (1998). Gaze-centered remapping of remembered visual space in an open-loop pointing task. *Journal of Neuroscience*, *18*(4), 1583-1594.

Jeannerod, M. (1984). The timing of natural prehension movements. *Journal of Motor Behavior*, 16, 235–254.

Knapen, T., Rolfs, M., & Cavanagh, P. (2009). The reference frame of the motion aftereffect is retinotopic. *Journal of Vision*, *9*(5), 16-16.

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, J. J. (1986). Optic flow. *Vision Research*, *26*(1), 161-179.

Koenderink, J. J., & Van Doorn, A. J. (1991). Affine structure from motion. *Journal of the Optical Society of America, A*, *8*(2), 377-385.

Koenderink, J. J., van Doorn, A. J., Kappers, A. M., & Lappin, J. S. (2002). Large-scale visual frontoparallels under full-cue conditions. *Perception*, *31*(12), 1467-1476.

Kothuri, R., Ravada, S., Sharma, J., & Banerjee, J. (2003). *U.S. Patent No. 6,505,205*. Washington, DC: U.S. Patent and Trademark Office.

Land, M. F., & McLeod, P. (2000). From eye movements to actions: how batsmen hit the ball. *Nature Neuroscience*, 3(12), 1340-1345.

Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, *28*(11), 1311-1328.

Lawrence, G. P., Khan, M. A., Buckolz, E., & Oldham, A. R. (2006). The contribution of peripheral and central vision in the control of movement amplitude. *Human Movement Science*, *25*(3), 326-338.

Lee, D. N., & Reddish, P. E. (1981). Plummeting gannets: a paradigm of ecological optics. *Nature*, *293*(5830), 293-294.

MacKay, D. M. (1973). Visual stability and voluntary eye movements. In *Central Processing of Visual Information A: Integrative Functions and Comparative Data* (pp. 307-331). Springer Berlin Heidelberg.

Makin, T. R., Holmes, N. P., Brozzoli, C., & Farnè, A. (2012). Keeping the world at hand: rapid visuomotor processing for hand–object interactions. *Experimental Brain Research*, 219(4), 421-428.

Marr; (1969) A theory of cerebellar cortex. *Journal of Physiol*ogy, 202 (1969), pp. 437–470

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

Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London B: Biological Sciences*, *207*(1167), 187-217.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B: Biological Sciences*, *200*(1140), 269-294.

Marr, D., & Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London B: Biological Sciences*, *204*(1156), 301-328.

Mayhew, J.E. (1973) After-effects of movement contingent on direction of gaze. *Vision Research*, 13(4), 877-880.
McBeath, M. K., Shaffer, D. M., & Kaiser, M. K. (1995). How baseball outfielders determine where to run to catch fly balls. *Science*, 268(5210), 569-573.
Melcher, D. (2005) Spatiotopic transfer of visual-form adaptation across saccadic eye movements. *Current Biology*, 15(19), 1745-1748.
Miall, R. C., Weir, D. J., Wolpert, D. M., & Stein, J. F. (1993). Is the cerebellum a smith predictor? *Journal of Motor Behavior*, 25(3), 203-216.
Mitchison, G. J., & McKee, S. P. (1987). The resolution of ambiguous stereoscopic matches by interpolation. *Vision Research*, *27*(2), 285-294.
Mitchison, G. J., & Westheimer, G. (1984). The perception of depth in simple figures. *Vision Research*, *24*(9), 1063-1073.
Müller, M., & Wehner, R. (1988). Path integration in desert ants, Cataglyphis fortis. *Proceedings of the National Academy of Sciences*, *85*(14), 5287-5290.
Newcombe, R. A., Lovegrove, S. J., & Davison, A. J. (2011). DTAM: Dense tracking and mapping in real-time. In IEEE *International Conference on Computer Vision* (pp. 2320-2327).
Ni, K., Kannan, A., Criminisi, A., & Winn, J. (2009). Epitomic location recognition. *Pattern Analysis and Machine Intelligence*, IEEE Transactions on, 31(12), 2158-2167.
Nishida, S. Y., Motoyoshi, I., Andersen, R. A., & Shimojo, S. (2003). Gaze modulation of visual aftereffects. *Vision Research*, *43*(6), 639-649.
O'Keefe, J. (1979). A review of the hippocampal place cells. *Progress in Neurobiology*, *13*(4), 419-439.
O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map* (Vol. 3, pp. 483-484). Oxford: Clarendon Press.
O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(05), 939-973.
Petrov, Y., & Glennerster, A. (2006). Disparity with respect to a local reference plane as a dominant cue for stereoscopic depth relief. *Vision Research*, 46(26), 4321-4332.
Poucet, B., Sargolini, F., Song, E. Y., Hangya, B., Fox, S., & Muller, R. U. (2014). Independence of landmark and self-motion-guided navigation: a different role for grid cells. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *369*(1635), 20130370
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.
Prince, S. J. D., Cumming, B. G., & Parker, A. J. (2002). Range and mechanism of encoding of horizontal disparity in macaque V1. *Journal of Neurophysiology*, *87*(1), 209-221.
Rav-Acha, A., Kohli, P., Rother, C., & Fitzgibbon, A. (2008). Unwrap mosaics: A new representation for video editing. In *ACM Transactions on Graphics (TOG)* (Vol. 27, No. 3, p. 17). ACM.
Sarlegna, F., Blouin, J., Vercher, J. L., Bresciani, J. P., Bourdin, C., & Gauthier, G. M. (2004). Online control of the direction of rapid reaching movements. *Experimental Brain Research*, *157*(4), 468-471.
Saunders, J. A., & Knill, D. C. (2004). Visual feedback control of hand movements. *Journal of Neuroscience*, *24*(13), 3223-3234.
Schnapp B., Warren W. (2007). Wormholes in virtual reality: What spatial knowledge is learned from navigation. *Journal of Vision*, 7 (9): 758, doi:10.1167/7.9.758
Servos, P., Goodale, M. A. & Jakobson, L. S. (1992). The role of binocular vision in prehension—A kinematic analysis. *Vision Research*, 32, 1513–1521
Smeets, J. B., Sousa, R., & Brenner, E. (2009). Illusions can warp visual space. *Perception*, *38*(10), 1467.
Svarverud, E., Gilson, S., & Glennerster, A. (2012). A demonstration of 'broken'visual space. PLoS One, *7*(3), e33782.
Tittle, J. S., Todd, J. T., Perotti, V. J., & Norman, J. F. (1995). Systematic distortion of perceived three-dimensional structure from motion and binocular stereopsis. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 663.
Torr, P. H., Szeliski, R., & Anandan, P. (2001). An integrated Bayesian approach to layer extraction from image sequences. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *23*(3), 297-303.
Triggs, B. (2000). Plane+parallax, tensors and factorization. In *Computer Vision-ECCV 2000* (pp. 522-538). Springer Berlin Heidelberg.
Triggs, B., McLauchlan, P. F., Hartley, R. I., & Fitzgibbon, A. W. (2000). Bundle adjustment—a modern synthesis. In *Vision algorithms: theory and practice* (pp. 298-372). Springer Berlin Heidelberg.
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.
Twinanda, A. P., Meilland, M., Sidibé, D., & Comport, A. I. (2013). On Keyframe Positioning for Pose Graphs Applied to Visual SLAM. *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, 5th Workshop on Planning, Perception and Navigation for Intelligent Vehicles*.
Warren, W.H., Rothman, D.B., Schnapp, B.H., & Ericson, J.D. (submitted) Wormholes in virtual space: From cognitive maps to cognitive graphs. Submitted to *Proceedings of the National Academy of Sciences*

Watt, R. J., & Morgan, M. J. (1985). A theory of the primitive spatial code in human vision. *Vision Research*, *25*(11), 1661-1674.
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.
Watt, S. J., & Bradshaw, M. F. (2003). The visual control of reaching and grasping: binocular disparity and motion parallax. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2), 404-415
Wehner, R., & Räber, F. (1979). Visual spatial memory in desert ants, Cataglyphis bicolor (Hymenoptera: Formicidae). *Experientia*, *35*(12), 1569-1571.
Whelan, T., Leutenegger, S., Salas-Moreno, R. F., Glocker, B., & Davison, A. J. (2015). ElasticFusion: Dense SLAM without a pose graph. In *Proceedings of Robotics: Science and Systems*.
Whitlock, J. R., Sutherland, R. J., Witter, M. P., Moser, M. B., & Moser, E. I. (2008). Navigating from hippocampus to parietal cortex. *Proceedings of the National Academy of Sciences*, *105*(39), 14755-14762.
Wilkie, R. M., Wann, J. P., & Allison, R. S. (2008). Active gaze, visual look-ahead, and locomotor control. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5), 1150-1164
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.