Navigating image space

Neuropsychologia 219 (2025) 109233

Contents lists available at ScienceDirect

Neuropsychologia

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

Navigating image space

Andrew Glennerster

School of Psychology and Clinical Language Sciences, University of Reading, Reading RG6 6AL, UK

ARTICLE INFO

Keywords:
Image space
Navigation
Fixation
Optic flow
Egocentric
Allocentric
3D
Spatial representation

ABSTRACT

Navigation means getting from here to there. Unfortunately, for biological navigation, there is no agreed definition of what we might mean by 'here' or 'there'. Computer vision ('Simultaneous Localisation and Mapping', SLAM) uses a 3D world-based coordinate frame but that is a poor model for biological spatial representation. Another possibility is to use an image-based rather than a map-based representation. The image-based strategy is made simpler if the observer maintains fixation on a stationary point in the scene as they move. This strategy would require a system for relating different fixation points to one another as the observer moves through the environment. I describe how this can be done by, first, relating fixations to an egocentric representation of visual direction and, second, encoding egocentric representations in a coarse-to-fine hierarchy. The coarsest level of this hierarchy is, in some sense, a world-based frame as it does not vary with eye rotation or observer translation. This representation could be implemented as a 'policy', a term used in reinforcement learning to describe a set of states and associated actions, or a 'graph' that describes how images or sensory states can be connected by actions. I discuss some of the psychophysical evidence relating to these differing hypotheses about spatial representation and navigation. I argue that this evidence supports image-based rather than map-based representation.

1. Moving through 3D space or image space

Navigation implies a representation of the observer's current location and of their goal plus some rules that will allow the observer to move from one to the other. The type of representation(s) that animals use remains a matter of debate. It does not have to be a 3D coordinate frame and many suggest that it is not (Mallot and Gillner, 2000; Warren, 2019; Meilinger et al., 2011; Parra-Barrero et al., 2023).

Fig. 1 illustrates two alternative types of approach to this problem. In Fig. 1(a), an observer walks along a path (O_1 to O_4) and records their location in a Cartesian, world-based frame of reference shown by the grid (Burgess et al., 2002; Byrne et al., 2007; Moser, 2014; Whittington et al., 2020). An alternative is shown in Fig. 1(b) where the observer takes the same path but now it also shows the points that the observer fixates (A, B, C) as they move. Thinking about the fixation point emphasises the retinal flow that the observer receives. Much of the ventral stream of visual processing is useful for identifying the fixated object while, conversely, the dorsal stream is relatively insensitive to the nature of the fixated object but instead provides highly sensitive information about the movement of the observer relative to the fixated object. This makes Fig. 1(b) a good starting point for thinking about the guidance of observer movement. We will explore this perspective in more detail in Section 4.

The literature covering hypotheses about retinal flow processing in the visual system is influenced by assumptions about the representations used for navigation, including the two alternatives sketched in Fig. 1. The underlying assumption in many models is that the visual system's goal is to recover the translation (movement in space) of the observer in a world-based frame like Fig. 1(a); if so, the argument goes, the visual system should decompose retinal flow into a 'translational' component and a 'rotational' component because the first of these can be used to recover the movement of the observer relative to the scene in a world-based frame of reference (Warren and Hannon, 1990; Heeger and Jepson, 1992; Lappe and Rauschecker, 1993; Shapiro et al., 1995; Matthis et al., 2022). The approach I describe here is different. I argue that the goal is not to recover the translation of the observer relative to a world-based frame but, instead, to change the current image into a goal image (similar to the approach used in current reinforcement learning algorithms for navigation (Zhu et al., 2017; Mirowski et al., 2018), i.e. the task is to navigate across a surface of images from the current image to the goal image. If this is the case, there is no need to decompose retinal flow into rotational and translational components (Glennerster et al., 2001; Glennerster, 2016).

In Section 3, I set out some of the problems that exist with the hypothesis of a map-based representation (Fig. 1(a)) then, in Sections 4–6, I outline an alternative approach based on navigating between the

E-mail address: a.glennerster@reading.ac.uk.

https://doi.org/10.1016/j.neuropsychologia.2025.109233
Received 11 January 2025; Received in revised form 26 May 2025; Accepted 16 July 2025
Available online 30 August 2025

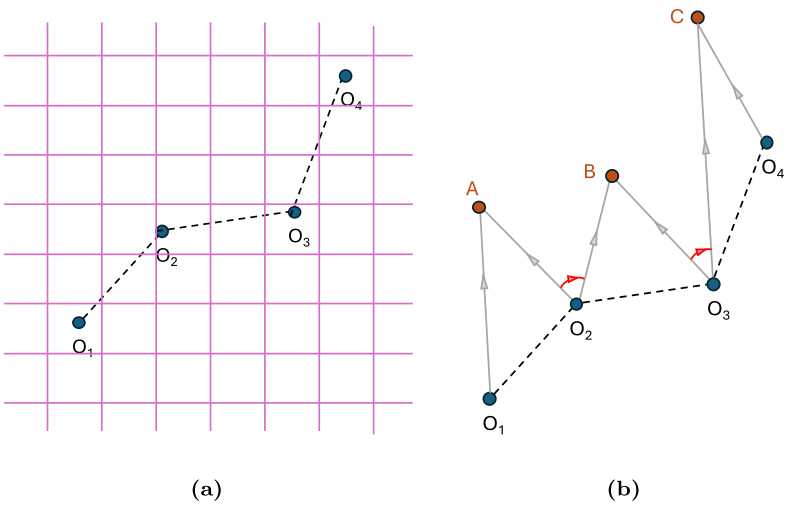

Fig. 1. Alternative methods of representing observer location. (a) Four locations of the observer (O_1 to O_4) are shown relative to a world centred reference frame (pink grid). (b) The same four locations are now shown including the fixation point for each segment of the movement ($O_1 \rightarrow O_2$, $O_2 \rightarrow O_3$ and $O_3 \rightarrow O_4$). The approach to representing the whole trajectory $O_1 \rightarrow O_4$ could be quite different from (a) if the first step is to estimate how the observer has moved in relation to the fixation point. This information would then need to be integrated across saccades (shown by red arrows). Fig. 3 describes one of these segments (e.g. $O_1 \rightarrow O_2$) in more detail.

current image and a goal image. First, in Section 2, I describe two anectodal examples from my own experience that illustrate why one might want to look for alternatives to the idea that we build a map of the world and use this to guide our actions.

2. Real world examples

Before going into details of, I describe two real-world examples of navigation that are difficult to explain if observers rely on a map of the environment to guide their actions. These provide motivation for thinking of alternatives to the idea that the visual system generates a world-based 3D model of the scene. The idea of a map is generally taken to mean that the visual system has an allocentric (world-based) representation (Tolman, 1948; O'Keefe and Nadel, 1978) that is rather like a 'survey' or birds-eye view of the scene. It does not need to be veridical, but it should be consistent across tasks.

The first real-world example relates to the disorientation that I sometimes observe when going down a spiral staircase. In the library I often work in, there is one that has quite a few 90° turns before I get to the lavatories in the basement and there are no windows on the way. By the time I reach the bottom I know with high confidence that I am facing either North, or West, or South or East rather than any intermediate orientation but I never know which of these is the case. If the representation that I use is a form of 3D reconstruction such as 'Simultaneous Localisation and Mapping' (SLAM) or a similar kind of map of the environment, that would not happen. But if the representation that I use is more like a set of images or neural states connected by actions, then this confusion is to be expected (and, incidentally, has no practical consequence because I still arrive at my goal). A set of images or states connected by actions is called a 'graph' where the images/states form the nodes and the edges joining the nodes are actions. A 'policy', $P(a|s)$, describes the actions, a, that are triggered by a set of states, s, so it is closely related to the idea of a graph of states connected by actions.

The second example is illustrated in Fig. 2. This shows a pair of lifts at the station in Reading on my way to work. The lifts can be entered from either side and, because they are built symmetrically, both entrances appear similar (lower images). When you emerge from the lift, your orientation is 180° different depending on the lift you were in, but the view is almost identical in either case (images outlined in green and red). For a long time, I would regularly come out of the lift and head off in the wrong direction. An inescapable conclusion, it seems to me, is that I was not simply using a map. If I built a map (like SLAM Davison, 2003) and then relied on this to guide my actions, it should would work equally well whether the scenes I saw coming out of the two lifts were similar or not. My confusion can only be explained if I was making image-dependent decisions about my next action, i.e. I was relying on a policy.

An elegant demonstration of a very similar conclusion comes from a study in which participants are asked to mime the action of driving a car including changing lane on a motorway. Almost no-one can do this (instead, they turn the wheel one way then back to the centre, which in real life would result in them veering off the road) (Wallis et al., 2007). The conclusion is, as with the example of emerging from the station lifts, that people do not form a map of the world and make a motor plan that would be appropriate for that map. Instead, they follow a policy, with every action followed by an image that triggers the next action and, for most tasks, this is sufficient to accomplish their goals.

3. Moving relative to a world-based map

The majority of physiologically-based models of navigation include a world-based representation of space and heading direction that are presumed to be dependent on the hippocampus and entorhinal cortex (O'Keefe et al., 1998; Howard et al., 2014). The idea of an allocentric or 'world-based' representation consists of two elements. One is a representation of the location of *objects*, which relies on 3D transformations of sensory information from egocentric to allocentric coordinate frames (Section 3.1). The other is a representation of the location of *the observer* in a world-based frame of reference (Section 3.2).

3.1. 3D coordinate transformation

If the visual system generates true 3D representations of the scene, e.g. in visual cortex, and these are used to contribute to a 3D world-based representation as the observer moves around, then somewhere in the brain there must be a process of coordinate transformation between these different reference frames. The posterior parietal cortex has long been associated with this hypothetical operation (e.g. retinal, head-, hand-, body- or world-centred coordinates) (Andersen and Zipser, 1988; Andersen et al., 1997; Denève and Pouget, 2003). The neural mechanisms that have been proposed to date are very complex. One

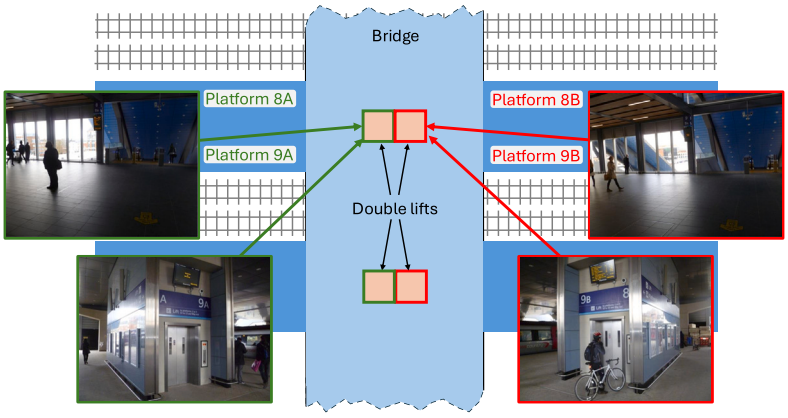

Fig. 2. Image-based navigation. The lifts from the platform to the bridge in Reading station are symmetrical (outlined in green and red here) which means that the view when you go in at platform level is similar whichever side you go in (lower red and green images) and the view when you exit the lift at the bridge level is almost *identical* (compare red and green upper images). Of course, the direction you face in each case as you emerge from the lift is 180° different. This led me to take the wrong turn about 50% of the time on my way to work.

example involves duplicating a representation many times in slightly different coordinate frames then 'gating' the output so that only one of the candidate representations is copied to the next stage of processing (Byrne et al., 2007). The logic applied here would get very much more complex if it were to be extended to the case of translation (movement in space) of the observer because there are so many more options for translation and it has no obvious limit. There are no detailed proposals for a neural mechanism to implement transformations of this type.

Some papers that discuss neural mechanisms for carrying out coordinate transformations suggest possible simplifications. One option is to update the coordinates of only a single object in the scene instead of the whole scene or to update a single difference vector (e.g. between the hand and a target) (Medendorp et al., 2003; Smith and Crawford, 2001; Prime et al., 2011; Pouget et al., 2002; Blohm et al., 2009). These are more plausible as neural mechanisms, but they abandon the idea of a 3D coordinate transformation underlying a 3D world-based representation of the whole scene based on retinal input.

3.2. Using grid cells for navigation

A separate goal of the visual system is to identify the world-centred *location of the observer* (rather than other objects) as they move around. There is evidence that place cells in the hippocampus and grid cells in the entorhinal cortex are involved in encoding this information (Barry and Bush, 2012; Erdem and Hasselmo, 2012; Kubie and Fenton, 2012), although there is a debate about the extent to which these cells provide a regular 'grid-like' reference frame similar to a longitude–latitude coordinate frame. For example, an extreme claim, advocated and tested by Carpenter et al. (2015), is that a continuous grid-like pattern of responses might extend across two rooms that are separated by a corridor. If that were true, it would suggest that the brain could apply a coherent coordinate system across both rooms, just like an externally defined longitude–latitude system. This would be a remarkable finding and would not be predicted by the type of scheme advocated in this paper (Sections 4 to 6). The data that Carpenter et al. (2015) present in fact support a more modest assertion, namely that rats, once they have learned to distinguish and navigate successfully between the two rooms, develop a new pattern of grid cell firing in the second room once. Stronger evidence than this would be required to support the assertion that a single continuous grid-like coordinate frame encompassed both rooms, despite many citations to this effect and a clear illustration in the graphical abstract (Glennerster, 2023).

Taken at face value, grid cells provide a signal that is not at all like the grid pattern of an *Ordnance Survey* map or a longitude–latitude coordinate frame because the signals from these cells are entirely ambiguous. This is the opposite of the unique labelling system that allows navigation using a map. Bush et al. (2015) try to address this problem by suggesting an algorithm for interpreting grid cell firing rates (from 9 cells, three at each of three scales) to provide an estimate of signal of location. The proposed decoding is not straightforward.

A more recent and more plausible model for interpreting the output of neurons that fire in multiple spatial locations has been presented by Banino et al. (2018). In this modelling paper, the activity of 512 units (rather than 9 grid cells) is used to learn a policy (a mapping between sensory context an action, in this case pointing towards a rewarded target). The 512 units each have receptive fields that, like grid cells, occur in multiple locations in the environment but only a portion of them have a regular spacing like grid cells and this subset have no special status among the 512 contributing cells. There is no attempt in this algorithm to generate a map or anything like a longitude–latitude coordinate frame.

A different modelling approach is taken by Whittington et al. (2020) who describe a 'Tolman–Eichenbaum Machine' that relates different sensory states to one another by storing the actions that would take the agent from one sensory state to another. They describe this as a 'graph' of states where the nodes of the graph are states and the edges are actions that connect them. This is closely related to the idea of a policy, since every state has an associated action. The use of information from grid cells is different in this model compared to Banino et al. (2018). The grid cell output is explicitly linked to the visual input at each point in space and 'loop closure' (Mei et al., 2011) is rewarded during the learning which results in there being a metric structure to the graph. We will see a quite different approach to developing a metric representation of the scene, starting with a topological a graph, in Section 6.

4. Moving relative to a fixation point

In this section, we will examine in more detail the idea that much of the visual system is arranged to facilitate the movement of the observer relative to a fixated object. Then, in Section 5, we will examine the problem of knitting together multiple epochs of movements relative to *different* fixation points, i.e. the issue illustrated in Fig. 1(b).

Almost all animals adopt a pattern of movement in which they 'fixate and saccade'. A fly with eyes rigidly attached to its head and its head rigidly attached to its body will make saccades with its body and, in between saccades, it will fixate on an object as it moves. Essentially, whether the eyes are free to move in the head or relative to the body, what matters is the gaze. The head and body compensate for eye movements to ensure gaze is fixed on an object as the animal moves and then

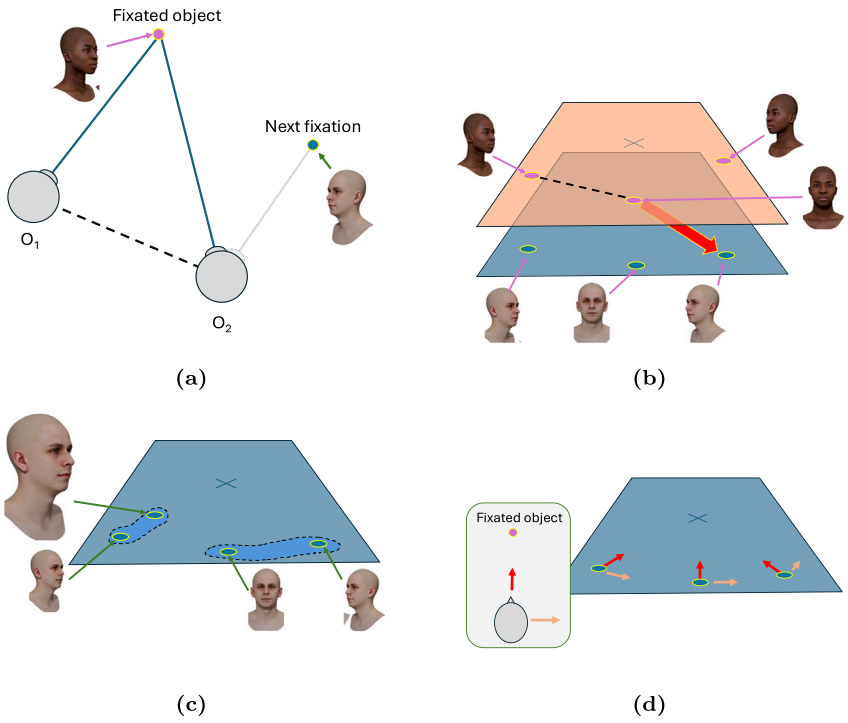

Fig. 3. Moving relative to a fixation point. (a) An observer moves from O_1 to O_2 while fixating on a point A and then makes a saccade to fixate object B, i.e. one transition from the trajectory illustrated in Fig. 1(b). (b) This shows a way to depict the retinal images that the observer receives during that movement. Every point on the pink surface represents an image and, for each image, the camera/eye is fixating the woman. The dashed line shows a path across this image space corresponding to the images that the observer would receive if they moved while fixating on the woman's head and then (red arrow) made a saccade to fixate the man (blue plain of images). (c) The responses of some types of neuron in the inferotemporal cortex can be plotted as a 'receptive field' on the surface of images, i.e. the set of images to which that neuron responds. The receptive field of a notional size-invariant neuron is shown on the left and of a view-invariant neuron on the right. (d) Dorsal stream neurons do not have receptive fields on the surface of images in the same way as ventral stream neurons. Instead, they signal motion across the surface, e.g. towards the fixated object (red arrows) or laterally (orange arrows).

makes a sudden switch to a new object (a saccade) (Land, 2009; Land and Nilsson, 2012). Gilchrist et al. (1997) show the same phenomenon in humans when extraocular fibrosis limits the ability to make saccades with the eyes: the head now makes saccade-like movements, so that the pattern of gaze movement is relatively unchanged despite the paralysis of the eye muscles.

Fig. 3 shows one of the epochs from Fig. 3(a). An observer moves from O_1 to O_2 while fixating point A and then makes a saccade to point B. Fig. 3(b) shows a different way to illustrate the same movement which highlights the image changes that are generated as the observer moves. Two surfaces are shown. One represents all the images that the camera/eye could see if it moved while maintaining fixation on a particular object. Each point on the surface corresponds to a different image. Nearby points on the surface correspond to images that could be obtained from nearby vantage points. Most animals, including humans, have a restricted pattern of eye movements that means they *must* maintain fixation for a period (e.g. about 200 ms) before they can make a saccade to a new object (i.e. jump to a new surface of images) (Land, 2009; Land and Nilsson, 2012). This is quite different from the general 6 degrees of freedom movement of a camera, which means that the retinal input to the visual system is radically different from the visual input to a computer vision algorithm (e.g. SLAM Davison, 2003). This leads to profound implications about the way that biological image processing is likely to take face compared to computer vision. In particular, it suggests that biological image processing to control navigation has a different goal from that of SLAM navigation systems.

The responses of neurons in the dorsal and ventral streams of visual processing are tailored to the restricted set of inputs that are produced by this pattern of eye movements. The ventral stream provides a rich source of information about which surface the current image is on (i.e. which object the observer is looking at) while the dorsal stream provides complementary information about the direction and magnitude of movement of the current image across the surface (largely independent of the object that the observer is looking at). Individual neurons in the ventral stream have responses that are relatively invariant to certain types of image change, e.g. the same stimulus viewed at different viewing distances (Ito et al., 1995) or from different viewing angles (Booth and Rolls, 1998) and this set of stimuli can be plotted as a region on the surface (Fig. 3(c)). It has been claimed that some neurons in macaque hippocampus respond to a very wide range of views of a location from different vantage points (Rolls et al., 1997). A hypothetical cell of this type would have a receptive field covering the entire surface shown in Fig. 3(c). All the way up the ventral stream, from complex cell in primary visual cortex to inferotemporal cortex to hippocampus, neurons in this stream tend to give a more stable output than a one might expect. Specifically, as the observer moves while fixating on an object, the retinal image changes. A neuron in inferotemporal cortex is relatively immune to these image changes.

The dorsal stream does the reverse and indicates change in the image caused by observer movement, relatively independent of the contents of the scene. It is not possible to draw the receptive fields of dorsal stream neurons on the surface of images (Fig. 3(c)) in the same way as for ventral stream neurons. Instead, their responses signal that the current image has moved across the surface in a particular direction. For example, Roy and Wurtz (1990) showed that neurons in MSTd respond to lateral head movement, while other MSTd cells respond to looming stimuli and self-motion in different directions with respect to the fixated object (Wild and Treue, 2021). The interpretation of these neurons as indicating a movement of the observer depends on the fact that the observer is fixating a point as they move. This pattern of fixational eye movement is maintained by a very fast feedback loop involving the nucleus of the optic tract (Ilg and Hoffmann, 1996).

Given that the observer *is* fixating, the movement of the observer relative to the fixation point can be decoded from a population of MSTd neurons (Lappe and Rauschecker, 1993; Gu et al., 2010).

If goals are defined in terms of desired images, then the role of visual processing in the dorsal stream could be much simpler than decomposing flow into rotational and translational components (Section 1). Specifically, if the goal of the observer is to change the current image into another image that is closer to (or at least on the path towards) the desired image, then the neurons in MSTd already encode the information in the relevant coordinate frame (a retinal one). They highlight the *change* in the image e.g. looming (approaching the fixated object) or the more complex type of image change that Roy and Wurtz (1990) describe (caused by moving sideways with respect to the fixated object). That is what is need in order to control the observer's movement in relation the fixated object.

It is worth remembering that the mechanisms for controlling movement suggested in this section are quite different from the method that would be involved in a general SLAM-like system of 3D reconstruction. In a fixating system, the rotation of the camera is yoked to translation so movement of the head leads to a unique image change. Essentially for rotation of the eye/camera, all 3 degrees of freedom are not free, they are determined by the translation of the eye in space (head movement) (Ferman et al., 1987; Glennerster et al., 2001). This makes it far simpler to control movement using image-based parameters than it would be if the rotation of the eye/camera was unconstrained (Fig. 1(a)).

In the next section, we consider translation (movement in space) of the observer over a larger scale and with multiple fixation targets.

5. An image-based frame of reference

This section explores the case illustrated in Fig. 1(b) where the observer moves while fixating on a series of different targets. The first step will be to consider how an egocentric representation of the visual direction of objects can be built up from a single vantage point. Then, I will discuss the consequences of observing a scene in which all the objects are distant. The egocentric representation in this case is independent of observer movement, so in one sense it is 'world-based'. Finally, I will show how a standard egocentric representation and this 'world-based' representation lie at two ends of a spectrum. This makes it possible to have a hierarchical, coarse-to-fine method of defining the address of the current image.

5.1. An egocentric representation of visual direction

Fig. 4 shows how, as an observer looks around, the images they receive can be put together into a single representation. This applies to a static monocular observer, so the eye is assumed to only rotate around its optic centre, not move in space. Fig. 4(a) shows the eye and two distant objects (mountain peaks). If the observer fixates one of these objects and then the other, the images from those two fixations can both contribute to a representation of the scene. This is shown in Fig. 4(b), when the eye is looking at Peak A, so an image of Peak A falls on the fovea. In Fig. 4(c), the eye has now rotated to look at Peak B so now the image of Peak B is on the fovea. The angle of rotation between A and B is α, i.e. this is the magnitude of the saccade that would take the eye from looking at A to looking at B. Fig. 4(d) shows how the retinal information from these two fixations can be represented in a common egocentric reference frame. It shows how the angles between different visible objects can be united in a common reference frame recording the relative visual directions of objects in the scene. For a more detailed account of this reference frame, see Glennerster et al. (2001).

Ultimately, it is possible to triangulate the entire sphere (assuming a transparent head and a freely-rotating eyeball) to record the relative visual direction of objects all around the observer. One might raise the objection that the number of pairs of points in a scene, and hence the number of potential saccades between them, is large (roughly n^2 for n points). However, there are ways to reduce the storage load and not include every possible pairing, for example by recording the location of fine scale features in the scene relative to coarse scale 'parent' features in a hierarchical structure (Watt, 1987, 1988; Koenderink and Van Doorn, 1991). There is psychophysical evidence to support such a hypothesis (Watt, 1987; Glennerster et al., 1996; Glennerster and McKee, 2004).

Although the egocentric reference frame in Fig. 4 is illustrated as a sphere, which is easy to grasp intuitively, an alternative implementation is to store a 'policy' as discussed in Section 3.2. A policy is a set of states where each state is associated with an action. In this case, the action is a saccade and the state combines information about the current and desired image, i.e. the image after the saccade. A list of these state–action-dyads makes a policy and, equivalently in this case, an egocentric representation.

5.2. A hierarchical address system for location

Next, we consider how this egocentric representation changes as the observer moves. Observer movement causes motion on the retina as objects at different distances move relative to one another (motion parallax). This information needs to be integrated and contribute to a representation that persists across eye movements (saccades). We will see that, as the observer moves, some elements of the representation in Section 5.1 change slowly while others change rapidly. This makes it possible to construct a hierarchical 'address', rather like a postal address (country, town, street, house), for defining spatial location.

Fig. 5 shows a scene that contains mountains, a forest in the middle distance and a picnic table close to the observer. We will see that the mountains provide a coarse scale 'address' (like 'UK' in a postal address) while the trees and the picnic table provide progressively finer detail about the observer's location. For example, Fig. 4 shows mountains that are very far away. If the observer translates (moves in space) by a few metres, the change in angles between the distant mountains will be so small that an observer cannot detect it. This means that the egocentric representation we have discussed in the previous section (Section 5.1) will be equally applicable wherever the observer moves their head (unless they walk a very long way). In this sense, the angles between distant mountains provide the coarsest possible component of the observer's current address.

The yellow plane drawn in the centre of Fig. 5 is related to the surfaces shown in Fig. 3 but is not identical. In Fig. 5, each point on the surface corresponds to an entire egocentric representation (like Fig. 4(d)) as viewed from one location in space. Fig. 3 was very similar in some ways, in that neighbouring points on the surface corresponded to the view from neighbouring locations in space but in that case the view was a single image. Now, in Fig. 5, the view corresponding to a point on the surface is a full 360 ° egocentric view of the scene.

If the scene only contains distant mountains, then egocentric representations that correspond to views from a wide range of locations are essentially indistinguishable. That is what is indicated by the large beige region in Fig. 5.

Just as a postal address can be refined by adding information about the town, street and house number, a visual address can be refined by adding in information about progressively closer objects in the scene. The picture on the right shows the same mountains but now with some trees that are closer to the observer. As the observer moves their head by a metre or two, there is detectable motion parallax between the trees and the mountains. This is shown in Fig. 5 by the fact that the light brown region is much smaller than the beige region, i.e. the region over which egocentric representations of the scene are indistinguishable is now much smaller. It is a subset of the larger region that was defined only by the having similar egocentric representations of the mountains. Finally, if the scene contains a picnic table close to the observer then the set of egocentric locations that are indistinguishable is even smaller

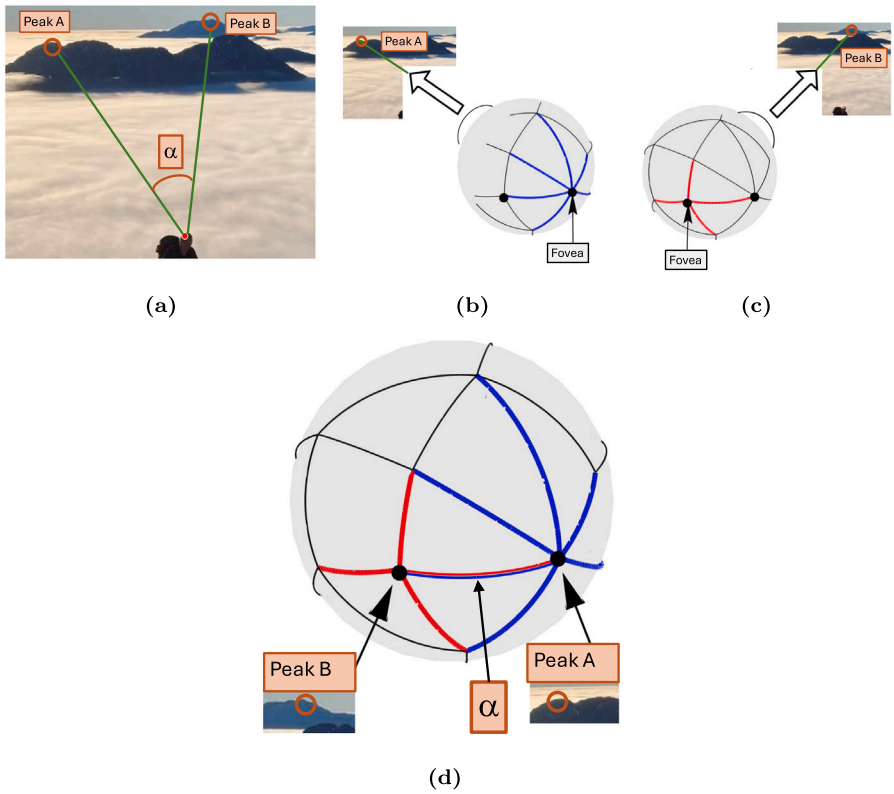

Fig. 4. An egocentric representation of visual direction. (a) This shows the angle, α, between two points measured at the eye. For distant points, this angle varies very little when the observer moves. (b) An eyeball with the fovea at the back, looking in the direction shown by the arrow. Here, the eye is looking at the mountain on the left (peak *A*). (c) Now the eye has rotated through an angle of α to look at the mountain on the right (peak *B*). (d) This shows information from (b) and (c) combined on a single sphere, i.e. independent of eye position. It shows the angles between pairs of points and hence the eye movement ('saccade') that would take the fovea from one object to another. In theory, this set of angles ('relative visual directions') can span the whole sphere and hence provide the information that would allow the eye to rotate from looking at any visible object in the scene to another. Hence, this is a type of egocentric representation of visual direction.

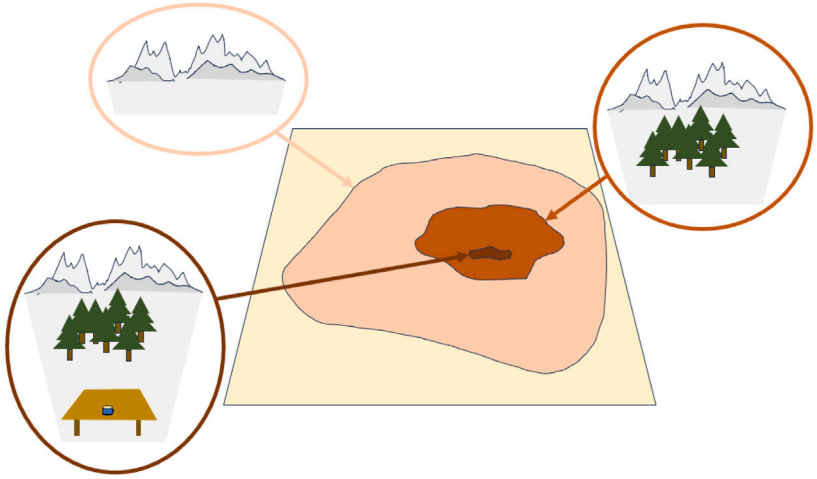

Fig. 5. A hierarchical address of an egocentric representation. (a) The yellow plane represents a surface of views in which each point corresponds to a full 360 degree egocentric view of the scene from one location (like Fig. 4(d)). Neighbouring points on this surface correspond to egocentric representations generated by viewing the scene from neighbouring locations in space. The beige region shows the set of egocentric representations that are (for practical purposes) indistinguishable if only distant objects are considered. When the angles between the trees and the mountains are included, the range of distinguishable egocentric representations narrows considerably (light brown region) and when nearby objects like the picnic table are included the range narrows even further (dark brown region). The egocentric representation that applies to the current location lies within all three regions; in other words, its address is hierarchical.

(dark brown region) and, again, forms a subset of the region defined only by the mountains and forest. The current egocentric view is defined by the objects visible from a single point in space and this egocentric representation corresponds to a single point on the surface in Fig. 5. That point lies within the dark brown region, which lies within the light brown region which lies within the beige region, just as Number 10 lies within Downing St and London and finally the UK. In other words, the address of the current egocentric representation has

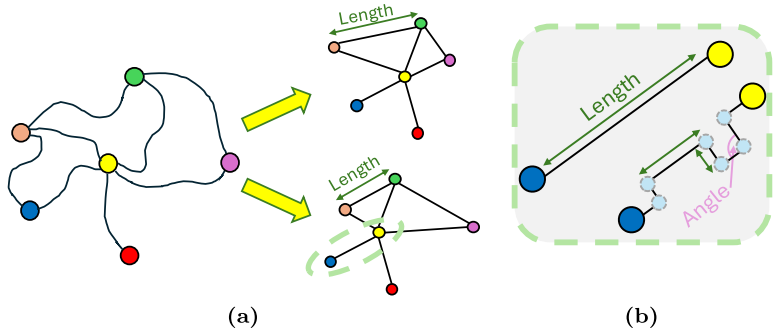

Fig. 6. A graph with edges described at different levels of detail. (a) A topological graph describes the connections (edges) between different locations (nodes) without any information about the length or angle of the edges. This is a coarse scale description of the layout in the sense that many different structures are compatible with the same topological structure. Two examples are shown here where, in each case, the length of the path between nodes is recorded in the graph. This level of detail allows planning of a shortest route to a goal. (b) An even finer scale of detail is shown here. The edge between the yellow and blue node can be described by its length or, as shown below, by a series of sub-turns each with an associated length and angle of rotation. In this way, sufficient detail can be added to the description of each edge that — in theory — the representation is impossible to distinguish from a metric map, at least in relation to the behaviour that it can support.

a hierarchical, scene-dependent address. A possible implementation of this hierarchical address system is described by Muryy et al. (2020).

We can now look back at Fig. 1 and see how it relates to the hierarchical description of location that we have described in this section. A path along a trajectory, O_1 to O_n, will change the observer's 'fine scale' address quite rapidly but the 'coarse scale' address will remain stable for a longer time as the observer moves. This is quite different from the idea that the observer recognises their location according to a fixed, 3D world-based coordinate frame (Fig. 1(a)).

This section has concentrated on the representation of location in an open space with objects visible both in the distance and nearby. Sometimes, this is called a 'vista space' (Meilinger et al., 2013). In the next section, we will look at evidence from participants navigating mazes where their view from any one location is restricted. Nevertheless, as we will see, these environments can also be described hierarchically and a policy is still a useful way to implement the representation.

6. Graphs for navigating mazes

Fig. 6 shows a number of places (shown by coloured circles) connected by routes. Knowing this topological map of connectivity between places would be sufficient to allow an observer to navigate between the different locations. This representation is a 'graph' with nodes (places) connected by edges (routes). In Fig. 6(a), the length and configuration of the routes is arbitrary, it simply indicates that a route exists between two locations so the spatial configuration of the places is not constrained. Siegel and White [22] suggest that observers start with a representation of connectivity like Fig. 6(a) and gradually add information about the edges between nodes. Two examples are shown on the right of Fig. 6(a), where each edge of the graph is now labelled with the length of the path between the two nodes it connects. This information allows an agent to travel by the shortest path to a goal whereas the topological information alone does not. The idea of adding more and more information about the edges is a powerful one. The information could be quite crude (e.g. 'shorter than average edge') but could be much more precise and include information about the length and curvature of the route and the angle between different routes that meet at a junction. Initially, these lengths and angles might be estimated quite crudely and that would make it impossible to unite all the nodes and edges together in a consistent 'map' of the environment. However, in theory, the lengths and angles of the routes connecting locations could be known with sufficient accuracy (lack of bias) and precision (lack of variability) that the representation becomes just like a map in the sense that the performance of an observer carrying out a range of tasks could be done equally well using a map or a highly calibrated graph.

Evidence in favour of this idea of a hierarchy of observers learning increasingly accurate graphs comes from experiments in virtual reality that allow experimenters to use physically impossible mazes and hence disentangle different hypotheses (Rothman and Warren, 2006; Warren et al., 2017; Warren, 2019). A similar approach was taken by Muryy and Glennerster (2018, 2021) who measured participants' accuracy for different tasks using a related virtual maze paradigm. In their case, the maze changed in configuration when the participant entered certain regions. The key result, as with Warren et al. (2017), was that the task matters, which should not be the case if participants rely on the same map to carry out different tasks such as pointing to an unseen target or finding the shortest route to a previously-visited location. If people instead rely on a hierarchy of labelled graphs and use the simplest one that they can for the task at hand then this apparently contradictory behaviour can be explained.

Chrastil and Warren (2014) review some of the evidence suggesting that participants use a hierarchy of tasks and, supporting these tasks, a corresponding hierarchy of representations. They describe survey knowledge (equivalent to building a map of the environment) as a different category of representation from labelled graphs, whereas Muryy and Glennerster (2021) advocate including the representation underlying survey knowledge under the same umbrella, as an extreme form of labelled graph (illustrated in Fig. 6(b)). The distinction may not be an important one.

The implementation of this hierarchy of representations could, as suggested in Sections 5.1 and 5.2, be as a policy. Fig. 6(a) is a coarse description of the layout of a scene. Adding information about the lengths of the paths between nodes in the graph refines the representation, so the policy (i.e. the contexts in which different actions are taken) becomes more constrained, i.e. the two situations on the right of Fig. 6(a), can be distinguished whereas before they could not. If the task is to go from the yellow to the green node by the shortest route, it is now possible to judge the relative lengths of the paths Yellow → Orange → Green versus Yellow → Purple → Green. In other words, a coarse scale representation of the action Yellow → Green has now been split onto two contexts with distinct actions associated with each.

The strength of this hierarchical approach is that it can explain why, in many situations, participants behave as if they are relying on a representation that is simpler than a Euclidean reconstruction of the scene. A good example of this logic, albeit not one from the navigation literature, comes from a paper by Glennerster et al. (1996) who asked participants to judge the shape (depth-to-height ratio) of a cylinder and also to compare the depths of cylinders at two distances. When participants were able to use a simple heuristic to do the task (compare cylinder depths) they were very accurate in their judgements. When they were forced to judge the shape (depth-to-height ratio) of a single

cylinder, they made large errors. This is very like the two examples we have just discussed of a hierarchical policy. In this case, the coarse scale representation might be sufficient to distinguish a concave versus a convex shape but with more information finer discriminations can be made, e.g. a triangular profile versus an elliptical one, with separate actions associated with each. Given even more information, the full Euclidean shape can be defined. Just as in the earlier examples of hierarchical encoding that we have discussed, the full Euclidean shape of an object is, according to this view, a subset or refinement of the coarser scale descriptions.

7. Discussion

In this paper, I have set out two opposing hypotheses about the reference frame that the visual system might use for navigation: graph-based (or a 'policy') versus a map-based representation. Some authors suggest that observers use graphs and maps 'interchangeably' (Peer et al., 2021). Instead, I have suggested that performance that seems to suggest a map-like representation may be an example of the way in which graph-based representations can be progressively refined (similar to the suggestion of Warren, 2019).

In the following sections, questions about whether a cognitive map is required for certain tasks that could not be carried out with a graph- or policy-based representation; how neurophysiologically inspired proposals are similar to or differ from computer vision approaches; how much visual processing is required for a policy representation and the role of other cues, such as proprioception in generating a spatial representation as a policy.

7.1. Path planning and other tasks that seem to require a map

Path planning is one example of a task that seems, at first sight, to be more difficult with a graph representation than using a 3D Cartesian coordinate representation (e.g. Kessler et al., 2024; Bush et al., 2015). However, suggestions have been made about ways to plan a path using a graph (e.g. Dai et al., 2020). There are also impressive demonstrations of navigational tasks including finding shortcuts that rely on a policy not a map (e.g. Banino et al., 2018). The way that human observers choose routes to a goal is one way to probe the type of representation they are using. For example, Muryy and Glennerster (2021) found that observers could plan a route successfully to a target in a distorted, physically-impossible maze while, at the same time, making large errors (up to 180 °) in pointing to targets. This dichotomy in performance on the two tasks is difficult to explain if both are based on the same internal map.

Similar results and conclusions are found by Warren and colleagues (Warren et al., 2017; Warren, 2019; Strickrodt et al., 2019). These experiments support the hypothesis that the visual system uses a range of heuristics, choosing different ones for different tasks. This can explain why a single underlying 3D representation provides such a poor account of human performance when participants carry out different tasks in the same environment (Glennerster et al., 1996; Svarverud et al., 2012; Muryy and Glennerster, 2021). One might think that there must be some tasks that could *only* be done using a map-like representation and would not be open to heuristics. However, given the argument that heuristics can be refined progressively (e.g. Fig. 6, Glennerster et al., 1996; Muryy and Glennerster, 2021; Glennerster, 2023), it may not be so easy to find an 'impossible' task for the graph model.

One argument that is often made against the idea of a graph of views as a spatial representation is that certain tasks should not be possible if the observer has not experienced the views in advance, e.g. predicting the view from the other side of a novel room. Aside from the fact that human observers turn out to be remarkably poor at this type of task (Vuong et al., 2019), there is modelling evidence that a 3D representation of the room is not required to carry out this task (Eslami et al., 2018). This shows that if a network is trained on a sufficient number of examples of similar environments, then one or two views of the novel room from one side of the room is sufficient to produce a remarkably accurate prediction of the view that would be seen from the other side of the room.

7.2. Rapid computation at 'runtime' versus large storage capacity

The two alternative approaches to spatial representation explored in this paper lead to very different challenges for models of neural implementation. If observers guide their movements using a world-based 3D reconstruction of the environment, then there needs to be a lot of computation at 'runtime'. One element of this is a decomposition of retinal flow into rotational and translational and rotational components (Lappe et al., 1999). Another is the rotation and translation of any egocentric representation into a world-based coordinate frame. In computer vision applications, these computations are carried out at frame rate, but detailed accounts are rare of the neurophysiological operations that could carry out equivalent transformations (Pouget et al., 2002; Byrne et al., 2007).

The alternative policy-based approach that I have advocated in this paper faces a quite different challenge. In some ways the problem is almost the converse. The computation at runtime could be far less than the 3D reconstruction approach, but the storage demand is much greater. The system must recognise a seemingly vast number of different situations (i.e. a particular image and a given task) and choose an action in response. The system must also compare the subsequent sensory input to the expected input and respond to any discrepancy between the two (Momennejad et al., 2017; Whittington et al., 2020). If the proposal is that all of these different situations are stored in advance, then any such model must face the challenge of explaining how they are learned and how they are stored.

There are many possible ways in which the problem of storage might be made manageable. One is the use of generalisation. For example, in the case of the movement shown in Fig. 1, it is not necessary to store *every* image that the observer could meet along the path. The visual system could store sufficient information to recognise the fixation targets A, B and C and then use a method that is independent of the nature of the fixation target to move relative to A, B or C. The signals from neurons in MSTd (Roy and Wurtz, 1990), which are largely blind to the nature of the fixated object, would be useful in this regard.

7.3. Fixation as a constraint

A central argument for the model presented here is that fixation is critical for the simplicity of motor control. One might ask whether the same is true for 3D construction algorithms, i.e. that yoking camera translation and rotation together and so reducing the number of degrees of freedom of camera movement could make 3D reconstruction simpler. This idea has been pursued by Aloimonos et al. (1987), Bandopadhay and Ballard (1990), Sandini and Tistarelli (2002), Daniilidis (1997). Daniilidis (1997) show explicitly how the computation of camera motion can be simplified when the camera fixates a scene point as it moves. However, unlike the policy-based proposal in this paper, the output in Daniilidis (1997) is still a 3D trajectory and a 3D description of the scene structure as in other photogrammetry algorithms.

7.4. Neurophysiological models that assume SLAM-like operations

It is important to realise how different computer vision approaches are compared to neurophysiological proposals in relation to 3D vision and navigation. Typically, neurophysiological accounts assume a two step process moving from image to egocentric representation (e.g. in posterior parietal cortex) followed by a transformation to world-based coordinates (e.g. in the hippocampus) (Andersen et al., 1997; Burgess et al., 1999; Savelli and Knierim, 2019). However, that is quite different

from the computer vision computation underlying SLAM ('photogrammetry') which finds the most likely structure of the scene and the most likely pose of the camera given a set of images of a static scene. There is no ego-centric intermediate stage in this process and hence no transformation from ego-centric to allocentric coordinates (Byrne et al., 2007). In this sense, the biological hypothesis is fundamentally different from SLAM. Also, a crucial output of SLAM is the world-based description of the scene structure. This is a quite different goal from generating outputs like a place cell (O'Keefe and Dostrovsky, 1971) or grid cell (Hafting et al., 2005) which signal the location of the observer not the world-based location of objects in the scene.

7.5. Visual control parameters

The discussion of navigation in this paper has been purely focused on vision, although it is extendible to other domains and to combinations of cues (Section 7.6). It is worth asking what type of visual processing is required if the representation is a graph- or policy-based one? The general answer is that the output of the processing should be sufficient to distinguish different contexts for action. If the task is to thread a needle, then the critical control parameter that is needed from moment to moment is a signal like the binocular disparity between the thread and the eye of the needle. The rest of the image is largely irrelevant to the task, although it is relevant to specifying the overall context (i.e. that the observer is threading a needle).

On the other hand, if the task is to move around an obstacle, then optic flow across the whole retina is important and, now, quite different aspects of the image are irrelevant to the task. When carrying out a complex sequence of actions, it is necessary to find the relevant visual information (and hence the appropriate set of neurons in the cortex) at the appropriate point in the sequence. Interestingly, Nienborg and Cumming (2014) have argued that a columnar organisation of the cortex is critical if the observer is to use sensory information to control an action in this way. Specifically, they suggest that the relevant cue (e.g. relative disparity) must be organised into cortical columns in order for those neurons to influence or reflect the choice of the animal (this is true of relative disparity in visual area MT but not in V1).

7.6. Idiothetic cues

Vision is not the only cue relevant to navigation. Other cues such as proprioception or knowledge of the interocular distance, collectively known as 'idiothetic cues', help the observer to infer their how they have moved and to estimate the size and shape of the scene. Indeed, there is evidence that the integration of visual and idiothetic cues can be close to optimal (Svarverud et al., 2010; Kang et al., 2023). Kang et al. (2023) present models of visual and idiothetic integration in rodents when environments are stretched, similar to the experiments and analysis by Svarverud et al. (2010). The idiothetic information is often assumed to be derived from grid cells (Hafting et al., 2005) although the problem of disambiguation of grid cell output (Bush et al., 2015) is not tackled in the Kang et al. (2023) paper. The predictions of optimal integration in Kang et al. (2023) provide a good fit to the experimental data (see also Nardini et al., 2008; Zhao and Warren, 2015; Chen et al., 2017 and modelling of these by Kessler et al., 2024).

A task can be learned with two cues present (e.g. visual and idiothetic) and then carried out with only one of the cues available. There is evidence that the different cues contribute to a common representation of the scene (e.g. Tcheang et al., 2011). This does not mean that the representation needs to be a 3-dimensional one. For example, a path through image space and a path through image+proprioceptive space can be closely related. The two cues can either support the same interpretation of the scene or be in conflict (Glennerster et al., 2009) but this type of evidence is not particularly helpful in discriminating between 3D and high dimensional coordinate frames.

The use of idiothetic cues to form a representation of space does not imply that the representation uses a 3D coordinate frame — that is a quite independent issue. One example that illustrates this independence is a paper by Banino et al. (2018) discussed in Section 3.2. Although grid cell activity is often cited as evidence of a 'map', this paper shows that an agent could learn to navigate and take short cuts without developing a Euclidean map (a 2- or 3-D representation with an origin and axes). Instead, the model used set of sensory contexts and actions that linked them, i.e. a 'policy'.

7.7. Visual ambiguity

If a representation is based on images rather than 3D structure, then it may be affected more by visual ambiguity — two scenes or viewpoints leading to the same image — than one based on 3D reconstruction. We saw that in the example of emerging from a lift in Section 2. Some image-based computer vision approaches to location suffer from mislocalisation if a camera image is ambiguous (Ni et al., 2009). Human observers suffer from errors in interpreting their motion and the motion of objects in the scene if the images are ambiguous, such as walking past a rotationally symmetric object (Wallach, 1987; Tcheang et al., 2005) or walking through an expanding room (Glennerster et al., 2006). However, observers are not totally lost, which shows that they can use proprioceptive cues and a history of the path they have been on (whether through 3D space or image space) to help disambiguate the location associated with the current image.

8. Conclusion

If 'here' and 'there' are defined in a space of images or neural states, then some of the more difficult challenges for finding plausible neurophysiological mechanisms disappear. One of these challenges is identifying how 3D coordinate transformations could be carried out, e.g. converting egocentric information to a world-based reference frame. However, observers still need some kind of reference frame and I have discussed how a reference frame for location might be constructed in vista spaces and in more visually constrained environments like mazes. In each case, the argument has been that the nervous system stores a 'policy', i.e. a set of states and an action associated with each state. I have focussed on the sensory aspect of the state, and in particular the image the observer receives, but a 'state' includes both sensory and task-related information (Section 5.1). Reinforcement learning is already using this type of approach to learn how to navigate (Section 1) and may be a valuable source of inspiration for understanding biological representations that can support navigation.

Acknowledgements

This research was supported by EPSRC/Dstl grant no. EP/N019423/1 and AHRC, United Kingdom grant no. AH/N006011/1.

Data availability

No data was used for the research described in the article.

References

Aloimonos, Y., Weiss, I., Bandopadhay, A., 1987. Active vision. In: International Conference on Computer Vision. pp. 35–54.

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. Annu. Rev. Neurosci. 20 (1), 303–330.

Andersen, R.A., Zipser, D., 1988. The role of the posterior parietal cortex in coordinate transformations for visual–motor integration. Can. J. Physiol. Pharmacol. 66 (4), 488–501.

Bandopadhay, A., Ballard, D.H., 1990. Egomotion perception using visual tracking. Comput. Intell. 7, 39–47.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M.J., Degris, T., Modayil, J., et al., 2018. Vector-based navigation using grid-like representations in artificial agents. Nature 557 (7705), 429–433.
Barry, C., Bush, D., 2012. From A to Z: a potential role for grid cells in spatial navigation. Neural Syst. & Circuits 2, 1–8.
Blohm, G., Keith, G.P., Crawford, J.D., 2009. Decoding the cortical transformations for visually guided reaching in 3D space. Cerebral Cortex 19 (6), 1372–1393.
Booth, M.C.A., Rolls, E.T., 1998. View-invariant representations of familiar objects by neurons in the inferior temporal cortex. Cerebral Cortex 8, 510–525.
Burgess, N., Jeffery, K.J., O'Keefe, J., 1999. The Hippocampal and Parietal Foundations of Spatial Cognition. OUP, Oxford.
Burgess, N., Maguire, E.A., O'Keefe, J., 2002. The human hippocampus and spatial and episodic memory. Neuron 35 (4), 625–641.
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. Psychol. Rev. 114 (2), 340.
Carpenter, F., Manson, D., Jeffery, K., Burgess, N., Barry, C., 2015. Grid cells form a global representation of connected environments. Curr. Biol. 25 (9), 1176–1182.
Chen, X., McNamara, T.P., Kelly, J.W., Wolbers, T., 2017. Cue combination in human spatial navigation. Cogn. Psychol. 95, 105–144.
Chrastil, E.R., Warren, W.H., 2014. From cognitive maps to cognitive graphs. PLoS One 9 (11), e112544.
Dai, T., Zheng, W., Sun, J., Ji, C., Zhou, T., Li, M., Hu, W., Yu, Z., 2020. Continuous route planning over a dynamic graph in real-time. Procedia Comput. Sci. 174, 111–114.
Daniilidis, K., 1997. Fixation simplifies 3D motion estimation. Comput. Vis. Image Underst. 68 (2), 158–169.
Davison, A.J., 2003. Real-time simultaneous localisation and mapping with a single camera. In: ICCV. pp. 1403–1410.
Denève, S., Pouget, A., 2003. Basis functions for object-centered representations. Neuron 37 (2), 347–359.
Erdem, U.M., Hasselmo, M., 2012. A goal-directed spatial navigation model using forward trajectory planning based on grid cells. Eur. J. Neurosci. 35 (6), 916–931.
Eslami, S.A., Rezende, D.J., Besse, F., Viola, F., Morcos, A.S., Garnelo, M., Ruderman, A., Rusu, A.A., Danihelka, I., Gregor, K., et al., 2018. Neural scene representation and rendering. Science 360 (6394), 1204–1210.
Ferman, L., Collewijn, H., Jansen, T., Van den Berg, A., 1987. Human gaze stability in the horizontal, vertical and torsional direction during voluntary head movements, evaluated with a three-dimensional scleral induction coil technique. Vis. Res. 27 (5), 811–828.
Gilchrist, I.D., Brown, V., Findlay, J.M., 1997. Saccades without eye movements. Nature 390 (6656), 130–131.
Glennerster, A., 2016. A moving observer in a three-dimensional world. Phil. Trans. R. Soc. B 371 (1697), 20150265.
Glennerster, A., 2023. Understanding 3D vision as a policy network. Philos. Trans. R. Soc. B 378 (1869), 20210448.
Glennerster, A., Hansard, M.E., Fitzgibbon, A.W., 2001. Fixation could simplify, not complicate, the interpretation of retinal flow. Vis. Res. 41, 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: International Dagstuhl Seminar, Dagstuhl Castle, Germany, July 13-18, 2008. Revised Papers. Springer, pp. 193–208.
Glennerster, A., McKee, S., 2004. Sensitivity to depth relief on slanted surfaces. J. Vis. 4 (5), 3, URL https://doi.org/10.1167/4.5.3.
Glennerster, A., Rogers, B.J., Bradshaw, M.F., 1996. Stereoscopic depth constancy depends on the subject's task. Vis. Res. 36, 3441–3456.
Glennerster, A., Tcheang, L., Gilson, S.J., Fitzgibbon, A.W., Parker, A.J., 2006. Humans ignore motion and stereo cues in favour of a fictional stable world. Curr. Biol. 16, 428–443.
Gu, Y., Fetsch, C.R., Adeyemo, B., DeAngelis, G.C., Angelaki, D.E., 2010. Decoding of MSTd population activity accounts for variations in the precision of heading perception. Neuron 66 (4), 596–609.
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.
Heeger, D.J., Jepson, A.D., 1992. Subspace methods for recovering rigid motion I: Algorithm and implementation. Int. J. Comput. Vis. 7, 95–117.
Howard, L.R., Javadi, A.H., Yu, Y., Mill, R.D., Morrison, L.C., Knight, R., Loftus, M.M., Staskute, L., Spiers, H.J., 2014. The hippocampus and entorhinal cortex encode the path and euclidean distances to goals during navigation. Curr. Biol. 24 (12), 1331–1340.
Ilg, U.J., Hoffmann, K.-P., 1996. Responses of neurons of the nucleus of the optic tract and the dorsal terminal nucleus of the accessory optic tract in the awake monkey. Eur. J. Neurosci. 8 (1), 92–105.
Ito, M., Tamura, H., Fujita, I., Tanaka, K., 1995. Size and position invariance of neuronal responses in monkey inferotemporal cortex. J. Neurophysiol. 73 (1), 218–226.
Kang, Y., Wolpert, D., M., L., 2023. Spatial uncertainty and environmental geometry in navigation. BioRxiv (Preprint). URL https://10.1101/2023.01.30.526278.
Kessler, F., Frankenstein, J., Rothkopf, C.A., 2024. Human navigation strategies and their errors result from dynamic interactions of spatial uncertainties. Nat. Commun. 15 (1), 5677.
Koenderink, J.J., Van Doorn, A.J., 1991. Affine structure from motion. J. Opt. Soc. Am. A 8 (2), 377–385.
Kubie, J.L., Fenton, A.A., 2012. Linear look-ahead in conjunctive cells: an entorhinal mechanism for vector-based navigation. Front. Neural Circuits 6, 20.
Land, M.F., 2009. Vision, eye movements, and natural behavior. Vis. Neurosci. 26 (1), 51–62.
Land, M.F., Nilsson, D.-E., 2012. Animal Eyes. OUP Oxford.
Lappe, M., Bremmer, F., van den Berg, A.V., 1999. Perception of self-motion from visual flow. Trends Cognit. Sci. 3 (9), 329–336.
Lappe, M., Rauschecker, J.P., 1993. A neural network for the processing of optic flow from ego-motion in man and higher mammals. Neural Comput. 5 (3), 374–391.
Mallot, H.A., Gillner, S., 2000. Route navigation without place recognition: what is recognized in recognition-triggered responses? Perception 29, 43–55.
Matthis, J.S., Muller, K.S., Bonnen, K.L., Hayhoe, M.M., 2022. Retinal optic flow during natural locomotion. PLoS Comput. Biol. 18 (2), e1009575.
Medendorp, W.P., Tweed, D.B., Crawford, J.D., 2003. Motion parallax is computed in the updating of human spatial memory. J. Neurosci. 23 (22), 8135–8142.
Mei, C., Sibley, G., Cummins, M., Newman, P., Reid, I., 2011. RSLAM: A system for large-scale mapping in constant-time using stereo. Int. J. Comput. Vis. 94, 198–214.
Meilinger, T., Riecke, B.E., Bülthoff, H.H., 2013. Local and global reference frames for environmental spaces. Q. J. Exp. Psychol. 1–28.
Meilinger, T., Wiener, J.M., Berthoz, A., 2011. The integration of spatial information across different perspectives. Mem. Cogn. 39, 1042–1054.
Mirowski, P., Grimes, M., Malinowski, M., Hermann, K.M., Anderson, K., Teplyashin, D., Simonyan, K., Zisserman, A., Hadsell, R., et al., 2018. Learning to navigate in cities without a map. In: Advances in Neural Information Processing Systems. pp. 2419–2430.
Momennejad, I., Russek, E.M., Cheong, J.H., Botvinick, M.M., Daw, N.D., Gershman, S.J., 2017. The successor representation in human reinforcement learning. Nature Human Behav. 1 (9), 680–692.
Moser, M.-B., 2014. Grid cells, place cells and memory. Nobel Lect. Available online at: http://www.nobelprize.org/nobel_prizes/medicine/laureates/2014/may-britt-moser-lecture.
Muryy, A., Glennerster, A., 2018. Pointing errors in non-metric virtual environments. In: Spatial Cognition XI: German Conference on Spatial Cognition. LNAI 11034. Springer, pp. 43–57.
Muryy, A., Glennerster, A., 2021. Route selection in non-euclidean virtual environments. PLoS One 16 (4), e0247818.
Muryy, A., Siddharth, N., Nardelli, N., Glennerster, A., Torr, P.H., 2020. Lessons from reinforcement learning for biological representations of space. Vis. Res. 174, 79–93.
Nardini, M., Jones, P., Bedford, R., Braddick, O., 2008. Development of cue integration in human navigation. Curr. Biol. 18 (9), 689–693.
Ni, K., Kannan, A., Criminisi, A., Winn, J., 2009. Epitomic location recognition. IEEE Trans. Pattern Anal. Mach. Intell. 31 (12), 2158–2167.
Nienborg, H., Cumming, B.G., 2014. Decision-related activity in sensory neurons may depend on the columnar architecture of cerebral cortex. J. Neurosci. 34 (10), 3579–3585.
O'Keefe, J., Burgess, N., Donnett, J.G., Jeffery, K.J., Maguire, E.A., 1998. Place cells, navigational accuracy, and the human hippocampus. Phil. Trans. R. Soc. B 353 (1373), 1333–1340.
O'Keefe, J., Dostrovsky, J., 1971. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain Res. 34, 171–175.
O'Keefe, J., Nadel, L., 1978. The Hippocampus as a Cognitive Map. Oxford University Press.
Parra-Barrero, E., Vijayabaskaran, S., Seabrook, E., Wiskott, L., Cheng, S., 2023. A map of spatial navigation for neuroscience. Neurosci. Biobehav. Rev. 152, 105200.
Peer, M., Brunec, I.K., Newcombe, N.S., Epstein, R.A., 2021. Structuring knowledge with cognitive maps and cognitive graphs. Trends Cogn. Sci. 25 (1), 37–54.
Pouget, A., Denève, S., Duhamel, J.-R., 2002. A computational perspective on the neural basis of multisensory spatial representations. Nature Rev. Neurosci. 3 (9), 741–747.
Prime, S.L., Vesia, M., Crawford, J.D., 2011. Cortical mechanisms for trans-saccadic memory and integration of multiple object features. Phil. Trans. R. Soc. B 366 (1564), 540–553.
Rolls, E.T., Robertson, R.G., Georges-François, P., 1997. Spatial view cells in the primate hippocampus. Eur. J. Neurosci. 9 (8), 1789–1794.
Rothman, D.B., Warren, W.H., 2006. Wormholes in virtual reality and the geometry of cognitive maps. J. Vis. 6 (6), 143. http://dx.doi.org/10.1167/6.6.143.
Roy, J.P., Wurtz, R.H., 1990. The role of disparity-sensitive cortical neurons in signalling the direction of self-motion. Nature 348, 160–162.
Sandini, G., Tistarelli, M., 2002. Active tracking strategy for monocular depth inference over multiple frames. IEEE Trans. Pattern Anal. Mach. Intell. 12 (1), 13–27.

Savelli, F., Knierim, J.J., 2019. Origin and role of path integration in the cognitive representations of the hippocampus: computational insights into open questions. J. Exp. Biol. 222 (Suppl. 1), jeb188912.
Shapiro, L.S., Zisserman, A., Brady, M., 1995. 3D motion recovery via affine epipolar geometry. Int. J. Comput. Vis. 16 (2), 147–182.
Smith, M.A., Crawford, J.D., 2001. Implications of ocular kinematics for the internal updating of visual space. J. Neurophysiol. 86 (4), 2112–2117.
Strickrodt, M., Bülthoff, H.H., Meilinger, T., 2019. Memory for navigable space is flexible and not restricted to exclusive local or global memory units. J. Exp. Psychol. [Learn. Mem. Cogn.] 45 (6), 993.
Svarverud, E., Gilson, S.J., Glennerster, A., 2010. Cue combination for 3D location judgements. J. Vis. 10 (1), 5. http://dx.doi.org/10.1167/10.1.5.
Svarverud, E., Gilson, S., Glennerster, A., 2012. A demonstration of 'broken' visual space. PLoS One (ISSN: 1932-6203) 7 (3), e33782. http://dx.doi.org/10.1371/journal.pone.0033782.
Tcheang, L., Bülthoff, H.H., Burgess, N., 2011. Visual influence on path integration in darkness indicates a multimodal representation of large-scale space. Proc. Natl. Acad. Sci. 108 (3), 1152–1157.
Tcheang, L., Gilson, S.J., Glennerster, A., 2005. Systematic distortions of perceptual stability investigated using immersive virtual reality. Vis. Res. 45 (16), 2177–2189.
Tolman, E.C., 1948. Cognitive maps in rats and men.. Psychol. Rev. 55 (4), 189.
Vuong, J., Fitzgibbon, A.W., Glennerster, A., 2019. No single, stable 3D representation can explain pointing biases in a spatial updating task. Sci. Rep. 9 (1), 1–13.
Wallach, H., 1987. Perceiving a stable environment when one moves. Annu. Rev. Psychol.: 1987 38, 1–27.
Wallis, G., Chatziastros, A., Tresilian, J., Tomasevic, N., 2007. The role of visual and nonvisual feedback in a vehicle steering task. J. Exp. Psychol. [Hum. Percept.] 33 (5), 1127.
Warren, W.H., 2019. Non-Euclidean navigation. J. Exp. Biol. 222 (Suppl 1), jeb187971.
Warren, W.H., Rothman, D.B., Schnapp, B.H., Ericson, J.D., 2017. Wormholes in virtual space: From cognitive maps to cognitive graphs. Cognition 166, 152–163.
Warren, Jr., W.H., Hannon, D.J., 1990. Eye movements and optical flow. J. Opt. Soc. Amer. A 7 (1), 160–169.
Watt, R.J., 1987. Scanning from coarse to fine spatial scales in the human visual system after the onset of a stimulus. J. Opt. Soc. Amer. A 4, 2006–2021.
Watt, R.J., 1988. Visual Processing: Computational, Psychophysical and Cognitive Research. Lawrence Erlbaum Associates, Hove.
Whittington, J.C., Muller, T.H., Mark, S., Chen, G., Barry, C., Burgess, N., Behrens, T.E., 2020. The Tolman-Eichenbaum Machine: unifying space and relational memory through generalization in the hippocampal formation. Cell 183 (5), 1249–1263.
Wild, B., Treue, S., 2021. Primate extrastriate cortical area MST: a gateway between sensation and cognition. J. Neurophysiol. 125 (5), 1851–1882.
Zhao, M., Warren, W.H., 2015. How you get there from here: Interaction of visual landmarks and path integration in human navigation. Psychol. Sci. 26 (6), 915–924.
Zhu, Y., Mottaghi, R., Kolve, E., Lim, J.J., Gupta, A., Fei-Fei, L., Farhadi, A., 2017. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In: 2017 IEEE international conference on robotics and automation (ICRA). IEEE, pp. 3357–3364.