University of Reading

Navigating image space

Article

Accepted Version

Creative Commons: Attribution-Noncommercial-No Derivative Works 4.0

Glennerster, A. ORCID: https://orcid.org/0000-0002-8674-2763 (2025) Navigating image space. Neuropsychologia, 219. 109233. ISSN 0028-3932 doi: 10.1016/j.neuropsychologia.2025.109233 Available at https://centaur.reading.ac.uk/124266/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.
Published version at: https://doi.org/10.1016/j.neuropsychologia.2025.109233
To link to this article DOI:
http://dx.doi.org/10.1016/j.neuropsychologia.2025.109233

Publisher: Elsevier

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Navigating image space

Andrew Glennerster
School of Psychology and Clinical Language Sciences
University of Reading, Reading RG6 6AL, UK
a.glennerster@reading.ac.uk

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 where the observer moves relative to a fixation point. This 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.

Keywords: Image space, navigation, fixation, optic flow, egocentric, allocentric, 3D, spatial representation.

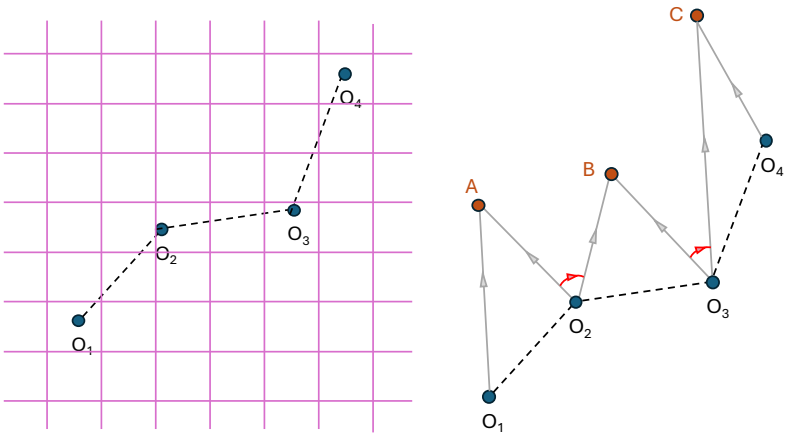
Manuscript submitted to *Neuropsychologia*: 26^{th} May 2025

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 [1–4].

Figure 1 illustrates two alternative types of approach to this problem. In Fig. 1a, 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 [5–8]. An alternative is shown in Fig. 1b 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. 1b 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. 1a; 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 [9–13]. 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 [14, 15]), 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 [16, 17].



(a) **(b)**

Figure 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 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.*

In Section 3, I set out some of the problems that exist with the hypothesis of a map-based representation (Fig. 1a) then, in Sections 4 to 6, I outline an alternative approach based on navigating between the 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 an alternative to a map-based representation, 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 [18, 19] 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', $P(\mathbf{a}|\mathbf{s})$, describes the actions, $\mathbf{a}$, that are triggered by a set of states, $\mathbf{s}$, so it is closely related to the idea of a graph of states

connected by actions.

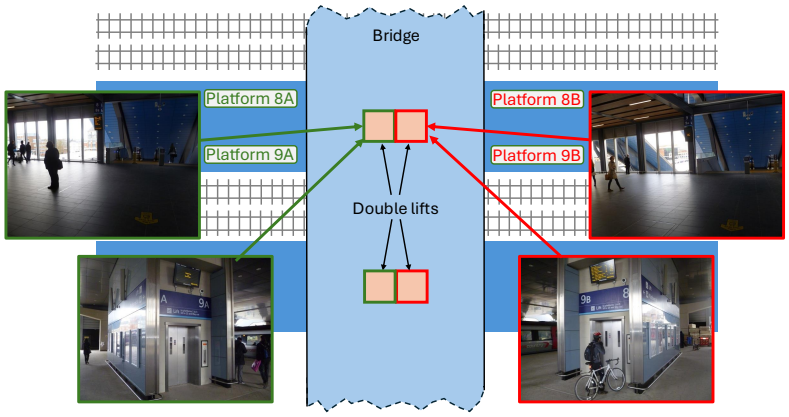


Figure 2: ***Image-based navigation.*** *The lifts from the platform to the bridge in Reading station are symmetrical (outlined in green and red here) which means that the view when you go in at platform level is similar whichever side you go in (lower red and green images) and the view when you exit 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. On my journey to work, unless I really concentrated, I would take the wrong turn about 50% of the time.*

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 [20]) and then relied on this to guide my actions, it should 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 correctly (instead, they turn the wheel one way then back to the centre, which in real life would result in them veering off the road) [21]. 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 [22, 23]. 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 a process of coordinate transformation must occur to transfer information 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) [24–26]. The neural mechanisms that have been proposed to date are very complex. One 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 [6]. The logic applied here would need to be very much more complex if it were to be extended to the case of translation (movement in space) of the observer. There are many more options for translation than for rotation and translation 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 tranformations 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) [27–31]. These are more plausible as neural mechanisms, but they abandon the idea of taking retinal inputs and using them to generate world-based representation of the whole scene.

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 the observer moves around. There is evidence that place cells in the hippocampus and grid cells in the entorhinal cortex are involved in encoding the location of the observer [32–34], although there is a debate about the extent to which these cells provide a regular 'grid-like' reference frame with a map-like role similar to a longitude-latitude coordinate frame. For example, an extreme claim, advocated and tested by Carpenter et al. [35], 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. [35] 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. Stronger evidence than this would be required to support the assertion that a single continuous grid-like coordinate frame encompassed both rooms (see [36]).

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. [37] 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. [38]. 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. [8] 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 [38]. The grid cell output is explicitly linked to the visual input at each point in space and 'loop closure' [39] is rewarded during the learning which results in there being a metric structure to the graph. We will see a discuss a different approach to developing a metric-like representation of the scene in Section 6.

4. Moving relative to a fixation point

In this section, we'll examine in more detail the idea that much of the visual system is arranged to facilitate the observer making a single transition, e.g. $O_1 \rightarrow O_2$ and then a saccade in Fig. 3a. Then, in Section 5, we'll examine the problem of knitting together multiple epochs of movements relative to *different* fixation points, i.e. the issue illustrated in Fig. 1b.

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 makes a sudden switch to a new object (a saccade) [40, 41]. Gilchrist et al. [42] 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.

Figure 3 shows one of the epochs from Fig. 3a. An observer moves from O_1 to O_2 while fixating point A and then makes a saccade to point B. Figure 3b 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) [40, 41]. 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 [20]). This leads to profound implications about the way that biological image processing is likely to proceed 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.

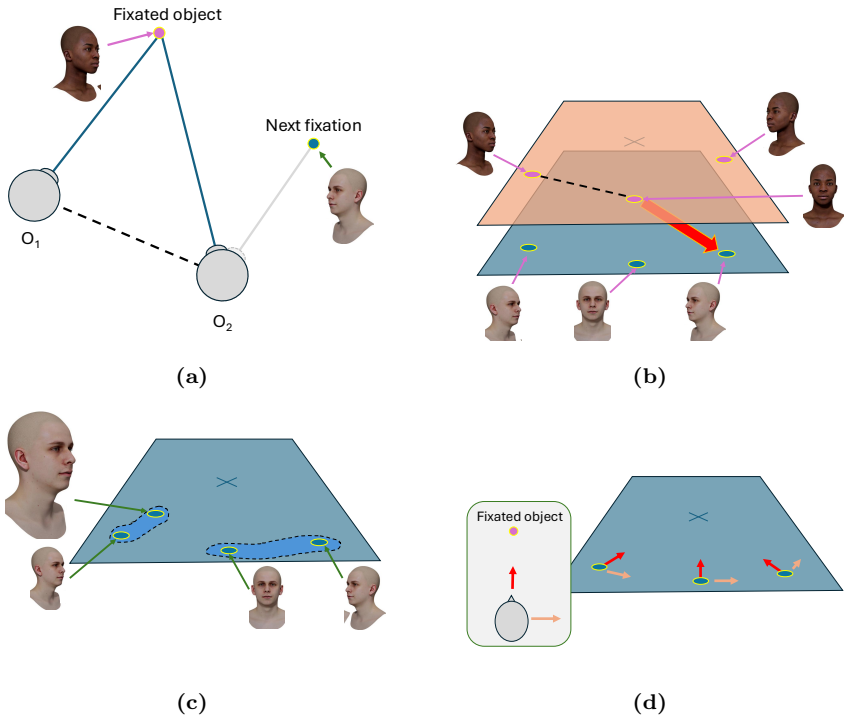


Figure 3: ***Moving relative to a fixation point****. (a) An observer moves from O_1 to O_2 while fixating on a point A and then makes a saccade to fixate object B, i.e. one transition from the trajectory illustrated in Fig. 1b. (b) This shows a way to depict the retinal images that the observer receives 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 in a particular direction, e.g. towards the fixated object (red arrows) or laterally (orange arrows).*

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 [43] or from different viewing angles [44] and this set of stimuli can be plotted as a region on the surface (Fig. 3c). 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 [45]. A hypothetical cell of this type would have a receptive field covering the entire surface shown in Fig. 3c. 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.

The dorsal stream does the reverse and indicates change 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. 3c) 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 [46] showed that neurons in the motion-sensitive cortical area 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 [47]. 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 [48]. 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 [11, 49].

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 [46] describe (caused by moving sideways with respect to the fixated object). This is the type of signal that is needed 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, all 3 degrees of freedom of rotation of the eye/camera are not actually free, they are determined by the translation of the eye in space (head movement) [16, 50]. 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. 1a).

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. 1b 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 of a world of distant points is independent of observer movement, so in one sense it is 'world-based'. Finally, I will show how a standard egocentric representation lies at one end of a spectrum, with a 'world-based' representation at the other end. This spectrum 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

Figure 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. Figure 4a 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. In Fig. 4b, the eye is looking at Peak A, so an image of Peak A falls on the fovea. In Fig. 4c, the eye has now rotated to look at Peak B so now the image of Peak B falls 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. Figure 4d shows how retinal information about 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 [16].

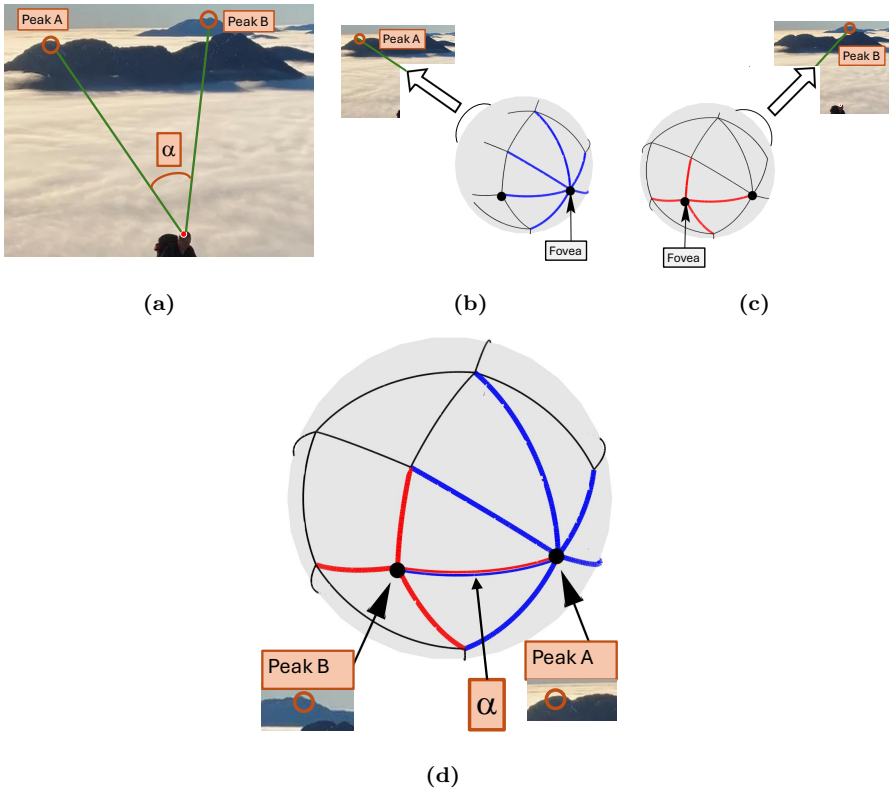


Figure 4: ***An egocentric representation of visual direction****. (a) This shows the angle, α, between two points measured at the eye. For distant points, this angle varies very little when the observer moves. (b) An eyeball with the fovea at the back, looking in the direction shown by the arrow. Here, 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 to look at any visible object in the scene. Hence, this is a type of egocentric representation of visual direction (but, at the same time, it is also a policy).*

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 [51–53]. There is psychophysical evidence to support such a hypothesis [51, 54, 55].

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'll 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.

Figure 5 shows a scene that contains mountains, a forest in the middle distance and a picnic table close to the observer. We'll 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. 4d) as viewed from one location in space. Figure 3 is very similar in some ways, in that neighbouring points in Fig. 3 correspond to the view from neighbouring locations in space, but in that case the 'view' was just 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.

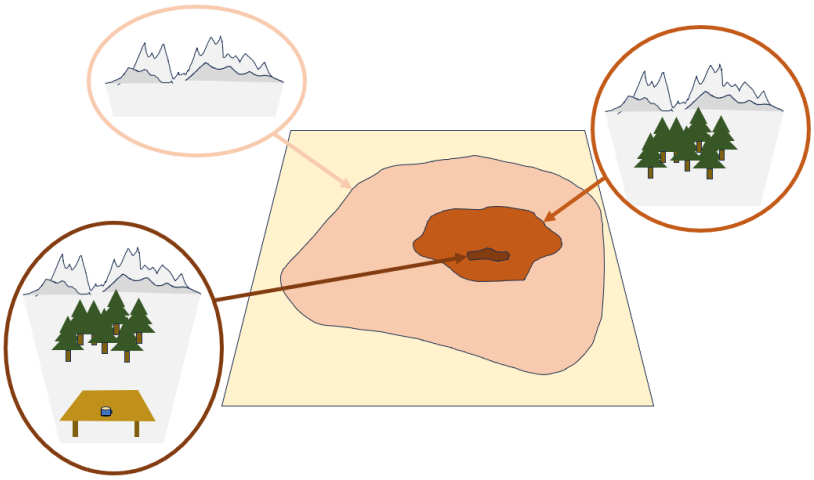


Figure 5: ***A hierarchical address of an egocentric representation.*** *(a) The yellow plane represents a surface of views in which each point corresponds to a full 360 degree egocentric view of the scene from one location (like Fig. 4d). Neighbouring points on this surface correspond to egocentric 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.*

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. Finally, if the scene contains a picnic table close to the observer then the set of egocentric locations that are indistinguishable is even smaller (dark brown region) and, again, this 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, which lies within London and finally the UK. In other words, the address of the current egocentric representation has a hierarchical, scene-dependent address. A possible implementation of this hierarchical address system is described by Muryy et al. [56].

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. 1a).

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' [57]. In the next section, we'll 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

Figure 6 shows a number of places (shown by coloured circles) connected by routes. This topological map of connectivity between places would be sufficient to allow an observer to navigate between the different locations. The representation is a 'graph' with nodes (places) connected by edges (routes). In Fig. 6a, the length and configuration of the routes is arbitrary, it simply indicates that a route exists between two locations so the spatial location of the places is not constrained. Siegel and White [22] suggest that observers start with a representation of connectivity like Fig. 6a and gradually add information about the edges between nodes. Two examples are shown on the right of Fig. 6a, 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, in which observers learning increasingly accurate graphs, comes from experiments in virtual reality that allow experimenters to use physically impossible mazes and hence disentangle different hypotheses [2, 58, 59]. A similar approach was taken by Muryy and Glennerster [60, 61] 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 [59], 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.

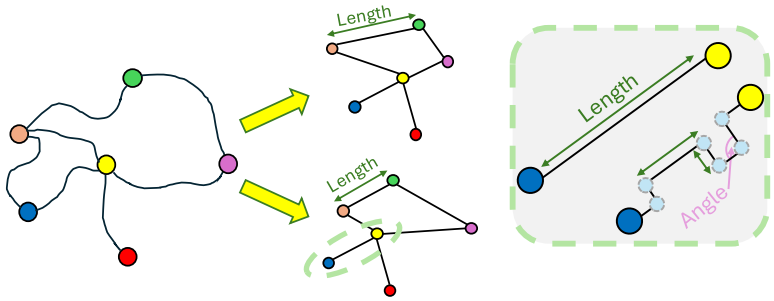


Figure 6: ***A graph with edges described at different levels of detail.*** *(a) A topological graph describes the connections (edges) between different locations (nodes) without any information about the length or angle of the edges. This is a coarse scale description of the layout in the sense that 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.*

Chrastil and Warren [62] 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 [61] advocate including the representation underlying survey knowledge under the same umbrella, i.e. as an extreme form of labelled graph (illustrated in Fig. 6b). The distinction between these interpretations may not be a critical

one, but the idea of a spectrum of increasingly sophisticated graphs is important.

The implementation of this hierarchy of representations could, as suggested in Sections 5.1 and 5.2, be as a policy. Figure 6a 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. This makes the policy (i.e. the contexts in which different actions are taken) more constrained. For example, the two situations on the right of Fig. 6a can be distinguished when extra information about turns is added, whereas before (pure topological graph in the left of Fig. 6a) 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 that have different 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. [54] 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 the case of shape discrimination, the coarse scale representation might only allow a distinction to be made between a concave and a convex shape. With more information, more detailed discriminations can be made, up to and including representing the true shape of the surface. The parallel with the navigation examples shown in Figs. 5 and 6 is that, in both these and the shape discrimination task, a true metric representation is at one end of a spectrum that includes, at the other end, far simpler categorisations of the stimulus.

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 (which is closely related to the idea of a 'policy') versus a map-based representation. Some authors suggest

that observers use graphs and maps 'interchangeably' [63]. Instead, I have argued, as others have [2], that graph-based representations can be progressively refined with the most highly calibrated graph having many of the functional features of a map (e.g. Fig. 6).

In the following sections, I discuss whether there are tasks that *could not* be carried out with a graph- or policy-based representation; how neurophysiologically-inspired proposals relate to computer vision approaches; what visual processing might be required for a policy representation; and the role of other cues, including proprioception, in generating a spatial representation.

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 it would be using a 3D Cartesian coordinate representation (e.g. [37, 64]). However, suggestions have been made about ways to plan a path using a graph (e.g. [65]). There are also impressive demonstrations of navigational tasks including finding shortcuts that rely on a policy not a map (e.g. [38]). 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 [61] 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 [2, 59, 66]. These experiments support the hypothesis that the visual system uses a range of heuristics, choosing different ones for different tasks. This can help to 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 [54, 61, 67]. 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, [36, 54, 61]), it may not be so easy to find an 'impossible' task for the graph model.

One version of the argument that certain tasks should not be possible using a graph-based representation concerns scenes that the observer has not experienced before, 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 [68], there is evidence from modelling studies that this task can be done without building a 3D map of the room. If a network is trained on a sufficient number of examples of similar environments, then two views of the novel room from one side of the room are sufficient to produce a remarkably accurate prediction of the view that would be seen from the other side of the room [69].

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 when it comes to possible 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 components [70]. 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 it is hard to see how similar operations could be carried out in the cortex. There are no detailed suggestions about how equivalent transformations could be carried implemented neurophysiologically [6, 30].

The alternative policy-based approach that I have advocated in this paper faces a quite different – and in some ways is almost the converse – challenge. The proposed computation at runtime is standard and familiar, 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 [8, 71]. If the proposal is that all of these different situations are stored in advance, then any such model must explain how so many could be learned and 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 identity of the fixation target to move relative to A, B or C. The signals from neurons in MSTd

([46], discussed in Section 4) are largely blind to the nature of the fixated object and 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 could make 3D reconstruction simpler. This idea has been pursued by [72–75]. Daniilidis [75] showed 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 [75] is still a camera trajectory within a 3D the scene and both are described in the same world-based coordinate frame.

7.4. Neurophysiological models versus SLAM

It is important to realise that neurophysiological proposals about 3D vision and navigation are radically different from standard computer vision approaches (like SLAM). 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) [25, 76, 77]. 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 [6]. 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 similar to those of a place cell [78] or grid cell [79] which signal the world-based location of the *observer* not the world-based location of *objects* in the scene.

7.5. Visual control parameters

The discussion so far has not addressed the type of visual processing that might be required for a graph-based representation. In general, the output of visual processing should be sufficient to distinguish different contexts for action. For example, if the

current task is to thread a needle, then the visual processing must generate a control parameter that is useful for that task, such as the binocular disparity between the thread and the needle. Of course, some visual processing is required to recognise the overall context (e.g. that the observer is in a room, threading a needle) but that does not change from moment to moment.

On the other hand, if the task is to move around an obstacle, then a very different type of visual processing becomes relevant for controlling the task. For example, in this case optic flow across the whole retina is important, quite unlike the needle-threading task. When carrying out a complex sequence of actions, the cortical task is to find the relevant set of neurons at the appropriate point in the sequence to control the observer's next movement.

Interestingly, Nienborg and Cumming [80] 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 sensory cue (e.g. relative disparity) is always organised into cortical columns in situations where experimenters find a tight correlation between neuronal firing rates and the animal's behavioural choice.

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 movement and the structure of the scene. There is evidence that the integration of visual and idiothetic cues can be close to optimal [81, 82]. Kang et al. [82] present models of visual and idiothetic integration in rodents when environments are stretched, similar to the experiments and analysis by Svarverud et al. [81]. The idiothetic information is often assumed (including in [82])to be derived from grid cells [79] although that postpones the problem of determining how grid cell firing indicates observer location (see [37]). The predictions of optimal integration in Kang et al. [82] provide a good fit to the experimental data (see also [83–85] and modelling by [64]).

There is evidence that visual and idiothetic cues contribute to a common representation of the scene (e.g. [86]) but this does not mean that the representation needs to be a 3-dimensional one. For example, if participants learn a path through a state

space in which states are made up of both visual and proprioceptive information, it can still be useful when only visual input or only proprioceptive input is available (i.e. the path still exists when projected onto a lower dimensional hyperplane).

7.7. Visual ambiguity

If a representation is based on images rather than 3D structure then there are situations in which it should be vulnerable to visual ambiguity whereas one based on 3D reconstruction would not be. This is exactly the problem that we discussed in Section 2 in relation to the visual ambiguity I experienced when emerging from a lift. The problem of visual ambiguity is recognised in some image-based computer vision approaches, where similar input images lead to mislocalisation [87]. If observers walk past a rotationally symmetric object, they have difficultly attributing the image changes correctly to either rotation of the object or movement of the observer [88, 89]. And if observers walk through an expanding room, the images that they receive (at least, when viewing the scene monocularly) are entirely ambiguous about the size of the room. In this case, idiothetic cues are very poor at resolving that ambiguity and portraying the correct size of the room [90]. This does not mean that images are the only determinant of scene structure. Idiothetic cues often play an important role, as we have discussed. Also, the path that the observer has taken to arrive at the current image is important (whether through 3D space or image space). This history often helps disambiguate the interpretation of location if the current input on its own is ambiguous.

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 grant no. AH/N006011/1.

References

[1] H. A. Mallot, S. Gillner, Route navigation without place recognition: what is recognized in recognition-triggered responses?, Perception 29 (2000) 43–55.

[2] W. H. Warren, Non-euclidean navigation, Journal of Experimental Biology 222 (2019) jeb187971.

[3] T. Meilinger, J. M. Wiener, A. Berthoz, The integration of spatial information across different perspectives., Memory & Cognition 39 (2011) 1042–1054.

[4] E. Parra-Barrero, S. Vijayabaskaran, E. Seabrook, L. Wiskott, S. Cheng, A map of spatial navigation for neuroscience, Neuroscience & Biobehavioral Reviews 152 (2023) 105200.

[5] N. Burgess, E. A. Maguire, J. O'Keefe, The human hippocampus and spatial and episodic memory, Neuron 35 (2002) 625–641.

[6] P. Byrne, S. Becker, N. Burgess, Remembering the past and imagining the future: a neural model of spatial memory and imagery., Psychological review 114 (2007) 340.

[7] M.-B. Moser, Grid cells, place cells and memory, Nobel Lecture. Available online at: http://www. nobelprize. org/nobel_prizes/medicine/laureates/2014/may-britt-moser-lecture. html (2014).

[8] J. C. Whittington, T. H. Muller, S. Mark, G. Chen, C. Barry, N. Burgess, T. E. Behrens, The Tolman-Eichenbaum Machine: unifying space and relational memory through generalization in the hippocampal formation, Cell 183 (2020) 1249–1263.

[9] W. H. Warren Jr, D. J. Hannon, Eye movements and optical flow, Journal of the Optical Society of America A 7 (1990) 160–169.

[10] D. J. Heeger, A. D. Jepson, Subspace methods for recovering rigid motion i: Algorithm and implementation, International Journal of Computer Vision 7 (1992) 95–117.

[11] M. Lappe, J. P. Rauschecker, A neural network for the processing of optic flow from ego-motion in man and higher mammals, Neural Computation 5 (1993) 374–391.

[12] L. Shapiro, A. Zisserman, J. M. Brady, 3D motion recovery via affine epipolar geometry 16 (1995) 147–182.

[13] J. S. Matthis, K. S. Muller, K. L. Bonnen, M. M. Hayhoe, Retinal optic flow during natural locomotion, PLOS Computational Biology 18 (2022) e1009575.

[14] Y. Zhu, D. Gordon, E. Kolve, D. Fox, L. Fei-Fei, A. Gupta, R. Mottaghi, A. Farhadi, Visual semantic planning using deep successor representations, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 483–492.

[15] P. Mirowski, M. Grimes, M. Malinowski, K. M. Hermann, K. Anderson, D. Teplyashin, K. Simonyan, A. Zisserman, R. Hadsell, et al., Learning to navigate in cities without a map, in: Advances in Neural Information Processing Systems, 2018, pp. 2419–2430.

[16] A. Glennerster, M. E. Hansard, A. W. Fitzgibbon, Fixation could simplify, not complicate, the interpretation of retinal flow, Vision Research 41 (2001) 815–834.

[17] A. Glennerster, A moving observer in a three-dimensional world, Phil. Trans. R. Soc. B 371 (2016) 20150265.

[18] E. C. Tolman, Cognitive maps in rats and men 55 (1948) 189—-208.

[19] J. O'Keefe, L. Nadel, The Hippocampus as a Cognitive Map, Oxford University Press, 1978.

[20] A. J. Davison, Real-time simultaneous localisation and mapping with a single camera, in: ICCV, 2003, pp. 1403–1410.

[21] G. Wallis, A. Chatziastros, J. Tresilian, N. Tomasevic, The role of visual and non-visual feedback in a vehicle steering task., Journal of Experimental Psychology: Human Perception and Performance 33 (2007) 1127.

[22] J. O'Keefe, N. Burgess, J. G. Donnett, K. J. Jeffery, E. A. Maguire, Place cells, navigational accuracy, and the human hippocampus, Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 353 (1998) 1333–1340.

[23] L. R. Howard, A. H. Javadi, Y. Yu, R. D. Mill, L. C. Morrison, R. Knight, M. M. Loftus, L. Staskute, H. J. Spiers, The hippocampus and entorhinal cortex encode the path and euclidean distances to goals during navigation, Current Biology 24 (2014) 1331–1340.

[24] R. A. Andersen, D. Zipser, The role of the posterior parietal cortex in coordinate transformations for visual–motor integration, Canadian journal of physiology and pharmacology 66 (1988) 488–501.

[25] R. A. Andersen, L. H. Snyder, D. C. Bradley, J. Xing, Multi-modal representation of space in the posterior parietal cortex and its use in planning movements, ARN 20 (1997) 303–330.

[26] S. Denève, A. Pouget, Basis functions for object-centered representations, Neuron 37 (2003) 347–359.

[27] W. P. Medendorp, D. B. Tweed, J. D. Crawford, Motion parallax is computed in the updating of human spatial memory, Journal of Neuroscience 23 (2003) 8135–8142.

[28] M. A. Smith, J. D. Crawford, Implications of ocular kinematics for the internal updating of visual space, Journal of Neurophysiology 86 (2001) 2112–2117.

[29] S. L. Prime, M. Vesia, J. D. Crawford, Cortical mechanisms for trans-saccadic memory and integration of multiple object features, Philosophical Transactions of the Royal Society B: Biological Sciences 366 (2011) 540–553.

[30] A. Pouget, S. Denève, J.-R. Duhamel, A computational perspective on the neural basis of multisensory spatial representations, Nature Reviews Neuroscience 3 (2002) 741–747.

[31] G. Blohm, G. P. Keith, J. D. Crawford, Decoding the cortical transformations for visually guided reaching in 3d space, Cerebral Cortex 19 (2009) 1372–1393.

[32] C. Barry, D. Bush, From a to z: a potential role for grid cells in spatial navigation, Neural systems & circuits 2 (2012) 1–8.

[33] U. M. Erdem, M. Hasselmo, A goal-directed spatial navigation model using forward trajectory planning based on grid cells, European Journal of Neuroscience 35 (2012) 916–931.

[34] J. L. Kubie, A. A. Fenton, Linear look-ahead in conjunctive cells: an entorhinal mechanism for vector-based navigation, Frontiers in neural circuits 6 (2012) 20.

[35] F. Carpenter, D. Manson, K. Jeffery, N. Burgess, C. Barry, Grid cells form a global representation of connected environments, Current Biology 25 (2015) 1176–1182.

[36] A. Glennerster, Understanding 3D vision as a policy network, Philosophical Transactions of the Royal Society B 378 (2023) 20210448.

[37] D. Bush, C. Barry, D. Manson, N. Burgess, Using grid cells for navigation, Neuron 87 (2015) 507–520.

[38] A. Banino, C. Barry, B. Uria, C. Blundell, T. Lillicrap, P. Mirowski, A. Pritzel, M. J. Chadwick, T. Degris, J. Modayil, et al., Vector-based navigation using grid-like representations in artificial agents, Nature 557 (2018) 429–433.

[39] C. Mei, G. Sibley, M. Cummins, P. Newman, I. Reid, Rslam: A system for large-scale mapping in constant-time using stereo, International journal of computer vision 94 (2011) 198–214.

[40] M. F. Land, Vision, eye movements, and natural behavior, Visual neuroscience 26 (2009) 51–62.

[41] M. F. Land, D.-E. Nilsson, Animal eyes, OUP Oxford, 2012.

[42] I. D. Gilchrist, V. Brown, J. M. Findlay, Saccades without eye movements 390 (1997) 130–131.

[43] M. Ito, H. Tamura, I. Fujita, K. Tanaka, Size and position invariance of neuronal responses in monkey inferotemporal cortex, Journal of neurophysiology 73 (1995) 218–226.

[44] M. C. A. Booth, E. T. Rolls, View-invariant representations of familiar objects by neurons in the inferior temporal cortex, Cerebral Cortex 8 (1998) 510–525.

[45] E. T. Rolls, R. G. Robertson, P. Georges-François, Spatial view cells in the primate hippocampus, European Journal of Neuroscience 9 (1997) 1789–1794.

[46] J. P. Roy, R. H. Wurtz, The role of disparity-sensitive cortical neurons in signalling the direction of self-motion, Nature 348 (1990) 160–162.

[47] B. Wild, S. Treue, Primate extrastriate cortical area mst: a gateway between sensation and cognition, Journal of neurophysiology 125 (2021) 1851–1882.

[48] U. J. Ilg, K.-P. Hoffmann, Responses of neurons of the nucleus of the optic tract and the dorsal terminal nucleus of the accessory optic tract in the awake monkey, European Journal of Neuroscience 8 (1996) 92–105.

[49] Y. Gu, C. R. Fetsch, B. Adeyemo, G. C. DeAngelis, D. E. Angelaki, Decoding of mstd population activity accounts for variations in the precision of heading perception, Neuron 66 (2010) 596–609.

[50] L. Ferman, H. Collewijn, T. Jansen, A. Van den Berg, Human gaze stability in the horizontal, vertical and torsional direction during voluntary head movements, evaluated with a three-dimensional scleral induction coil technique, Vision research 27 (1987) 811–828.

[51] R. J. Watt, 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 (1987) 2006–2021.

[52] R. J. Watt, Visual processing: computational, psychophysical and cognitive research, Lawrence Erlbaum Associates, Hove, 1988.

[53] J. J. Koenderink, A. J. van Doorn, Affine structure from motion 8 (1991) 377–385.

[54] A. Glennerster, B. J. Rogers, M. F. Bradshaw, Stereoscopic depth constancy depends on the subject's task, Vision Research 36 (1996) 3441–3456.

[55] A. Glennerster, S. P. McKee, Sensitivity to depth relief on slanted surfaces 4 (2004) 378–387.

[56] A. Muryy, N. Siddharth, N. Nardelli, A. Glennerster, P. H. Torr, Lessons from reinforcement learning for biological representations of space, Vision Research 174 (2020) 79–93.

[57] T. Meilinger, B. E. Riecke, H. H. Bülthoff, Local and global reference frames for environmental spaces, The Quarterly Journal of Experimental Psychology (2013) 1–28.

[58] D. B. Rothman, W. H. Warren, Wormholes in virtual reality and the geometry of cognitive maps, Journal of Vision 6 (2006) 143. doi:10.1167/6.6.143.

[59] W. H. Warren, D. B. Rothman, B. H. Schnapp, J. D. Ericson, Wormholes in virtual space: From cognitive maps to cognitive graphs, Cognition 166 (2017) 152–163.

[60] A. Muryy, A. Glennerster, Pointing errors in non-metric virtual environments, in: German conference on spatial cognition, Springer, 2018, pp. 43–57.

[61] A. Muryy, A. Glennerster, Route selection in non-euclidean virtual environments, PloS one 16 (2021) e0247818.

[62] E. R. Chrastil, W. H. Warren, From cognitive maps to cognitive graphs, PloS one 9 (2014) e112544.

[63] M. Peer, I. K. Brunec, N. S. Newcombe, R. A. Epstein, Structuring knowledge with cognitive maps and cognitive graphs, Trends in cognitive sciences 25 (2021) 37–54.

[64] F. Kessler, J. Frankenstein, C. A. Rothkopf, Human navigation strategies and their errors result from dynamic interactions of spatial uncertainties, Nature Communications 15 (2024) 5677.

[65] T. Dai, W. Zheng, J. Sun, C. Ji, T. Zhou, M. Li, W. Hu, Z. Yu, Continuous route planning over a dynamic graph in real-time, Procedia Computer Science 174 (2020) 111–114.

[66] M. Strickrodt, H. H. Bülthoff, T. Meilinger, Memory for navigable space is flexible and not restricted to exclusive local or global memory units., Journal of Experimental Psychology: Learning, Memory, and Cognition 45 (2019) 993.

[67] E. Svarverud, S. Gilson, A. Glennerster, A demonstration of 'broken' visual space., PLoS one 7 (2012) e33782. doi:10.1371/journal.pone.0033782.

[68] J. Vuong, A. W. Fitzgibbon, A. Glennerster, No single, stable 3d representation can explain pointing biases in a spatial updating task, Scientific reports 9 (2019) 1–13.

[69] S. A. Eslami, D. J. Rezende, F. Besse, F. Viola, A. S. Morcos, M. Garnelo, A. Ruderman, A. A. Rusu, I. Danihelka, K. Gregor, et al., Neural scene representation and rendering, Science 360 (2018) 1204–1210.

[70] M. Lappe, F. Bremmer, A. V. van den Berg, Perception of self-motion from visual flow, Trends in cognitive sciences 3 (1999) 329–336.

[71] I. Momennejad, E. M. Russek, J. H. Cheong, M. M. Botvinick, N. D. Daw, S. J. Gershman, The successor representation in human reinforcement learning, Nature human behaviour 1 (2017) 680–692.

[72] Y. Aloimonos, I. Weiss, A. Bandopadhay, Active vision, ICCV (1987) 35–54.

[73] A. Bandopadhay, D. H. Ballard, Egomotion perception using visual tracking, Computational Intelligence 7 (1990) 39–47.

[74] G. Sandini, M. Tistarelli, Active tracking strategy for monocular depth inference over multiple frames 12 (1990) 13–27.

[75] K. Daniilidis, Fixation simplifies 3D motion estimation, Computer Vision and Image Understanding 68 (1997) 158–169.

[76] N. Burgess, K. J. Jeffery, J. O'Keefe, The hippocampal and parietal foundations of spatial cognition, Oxford: OUP, 1999.

[77] F. Savelli, J. J. Knierim, Origin and role of path integration in the cognitive representations of the hippocampus: computational insights into open questions, Journal of Experimental Biology 222 (2019) jeb188912.

[78] J. O'Keefe, J. Dostrovsky, The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat, Brain Research 34 (1971) 171–175.

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

[80] H. Nienborg, B. G. Cumming, Decision-related activity in sensory neurons may depend on the columnar architecture of cerebral cortex, Journal of Neuroscience 34 (2014) 3579–3585.

[81] E. Svarverud, S. J. Gilson, A. Glennerster, Cue combination for 3d location judgements 10 (2010) 1–13. doi:http://dx.doi.org/10.1167/10.1.5.

[82] Y. Kang, D. Wolpert, L. M, Spatial uncertainty and environmental geometry in navigation, BioRxiv (preprint) (2023). URL: "https://10.1101/2023.01.30.526278".

[83] M. Nardini, P. Jones, R. Bedford, O. Braddick, Development of cue integration in human navigation, Current biology 18 (2008) 689–693.

[84] M. Zhao, W. H. Warren, How you get there from here: Interaction of visual landmarks and path integration in human navigation, Psychological science 26 (2015) 915–924.

[85] X. Chen, T. P. McNamara, J. W. Kelly, T. Wolbers, Cue combination in human spatial navigation, Cognitive Psychology 95 (2017) 105–144.

[86] L. Tcheang, H. H. Bülthoff, N. Burgess, Visual influence on path integration in darkness indicates a multimodal representation of large-scale space, Proceedings of the National Academy of Sciences 108 (2011) 1152–1157.

[87] K. Ni, A. Kannan, A. Criminisi, J. Winn, Epitomic location recognition, 2009.

[88] H. Wallach, Perceiving a stable environment when one moves 38 (1987) 1–27.

[89] L. Tcheang, S. Gilson, A. Glennerster, A. Parker, Perceiving a stable environment using immersive virtual reality. 31 (2002) S123.

[90] A. Glennerster, L. Tcheang, S. J. Gilson, A. W. Fitzgibbon, A. J. Parker, Humans ignore motion and stereo cues in favour of a fictional stable world 16 (2006) 428–43.